Learning Models that Match

Jacob Tyo

March 2024
CMU-ML-24-101

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Zachary C. Lipton (Chair, Carnegie Mellon University)
Jeff Schneider (Carnegie Mellon University)
James Hare (DEVCOM Army Research Laboratory)
Willie Neiswanger (University of Southern California)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Machine Learning.

Keywords: Machine Learning, Contrastive Learning, Authorship Identification, Meta-Learning, Motorcycle Racing Dataset, Multiple Instance Learning, Text Spotting, Person Search

To my parents, Frank and Terri Tyo, for their continual and unwavering love, support, and sacrifice throughout my life. Without them, I would not be the person I am today. To my wife, for being my steadfast companion and for pushing me forward when I was ready to give up. And to my son, Briggs, I hope this serves as a beacon of hope in your most trying endeavors.

Abstract

Contrastive learning has emerged as a critical methodology in machine learning applications, offering a pair-wise comparison perspective on data interpretation and model training. This thesis comprehensively examines contrastive learning models, emphasizing their development, application, and optimization for real-world scenarios. This thesis is structured into two main sections: the first explores practical applications in diverse domains such as authorship attribution, verification, and person re-identification, while the second focuses on methodological advancements aimed at enhancing model efficacy and adaptability.

In Part I, the thesis systematically evaluates the application of contrastive learning techniques across various fields, highlighting their strengths and limitations in real-world settings. Through detailed case studies, including the implementation of a photo-searching system for off-road motorcycle racing, this work assesses the adaptability and effectiveness of contrastive models under challenging conditions. The findings underscore the necessity for nuanced understanding and strategic application of these models to harness their full potential, especially concerning curating the right pairs during training.

Part II delves into developing innovative approaches to overcome the inherent challenges identified in contrastive learning. It introduces new algorithms and frameworks designed to refine the learning process, particularly in handling weakly labeled data and optimizing the influence of each sample on the overall loss (i.e. the pair curation). The proposed methodologies aim to bridge the gap between theoretical principles and practical utility, facilitating the creation of more robust, efficient, and versatile machine learning systems.

This thesis yields highly-performant authorship identification and person re-identification models, often achieving a new state-of-the-art. Furthermore, the insights drawn from analysis of these models and applications lead to the introduction of two methodologies that enhance model training. The first is a method for automatically adjusting the influence each data-point has on a model at a particular point in training, and the second method enables contrastive training among weakly labeled data via a contrastive extension to the multiple-instance learning framework. Together, these findings represent insight into the dynamics of contrastive learning, and present viable solutions to broaden their real-world applicability.

Acknowledgments

In high school, I was not particularly interested in academics, seeing them only as a means to an end for taking action. Nevertheless, the particularly talented, tenacious, tactful, and endlessly patient small-town teachers at Braxton County High–especially Charles Toumazos, Jill Lemon, Shirley Shuman, and Brenda Gibson–built the foundation that I skirted through undergrad with, chasing work experience over academics.

My options for entering the job market after my undergraduate degree left me yearning for more, pushing me towards further education. Roy Nutter Jr. saw potential in me, offering me a chance to dive into research as a master's student. This path led me to Katerina Goseva-Popstojanova, who welcomed me to her lab with a project to build an AI system for detecting software bugs in NASA code that can lead to security breaches. Despite my stubbornness and overconfidence, their mentorship and patience were unwavering, and for that, I am eternally grateful.

The pivotal moment in my academic career was my encounter with Nasser Nasrabadi during his neural networks class. He picked me out, and with his guidance and a consequential introduction to the Army Research Lab, set me on a trajectory that eventually pointed towards the need for a deeper research foundation. Tian Pham and Andrew Ladas were instrumental in this transition, advocating for my PhD pursuit, a gesture of support I will never forget. Since then, the support from Lance Kaplan, James Hare, and Carl Busart has been a consistent comfort.

The journey has been arduous, and now it is clear that it takes a village to raise a PhD. At the forefront was my advisor, Zachary Lipton. Navigating through my challenges, Zack's steadfast belief and strategic guidance were the beacons that kept me oriented toward my goals. His ability to discern when to apply pressure and when to give space was paramount to my growth, both professionally and personally. I also thank the rest of my committee, Jeff Schneider, James Hare, and Willie Neiswanger for their roles in my development as a researcher.

The initial years of the PhD were fraught with uncertainty, yet the guidance from Christoph Dann, Mariya Toneva, Maruan Al-Shedivat, Anthony Platanios, Avinava Dubey, Maria De-Arteaga, Biswajit Paria, and many others provided clarity and direction during these turbulent times. My officemates were pillars of support, offering respite and companionship through countless challenges.

The ACMI lab was more than just a workplace; it was a community. The discussions, feedback, and shared experiences, from lab dinners to frantic paper revisions, made the journey bearable and rewarding. To everyone in the lab and my PhD cohort, including Kundan Krishna, Helen Zhou, Leqi Liu, Saurabh Garg, Pratyush Maini, Shantanu Gupta, Tanya Marwah, Andrey Huang, Danish Pruthi, Divyansh Kaushik, and many more, your support and shared adventures have left an indelible mark on my heart.

The camaraderie and intellectual stimulation provided by my colleagues and academic collaborators–Bhuwan Dhingra, Benjamin Eysenbach, Youngseog Chung, Motolani Olarinre, and many others–enriched my PhD experience immeasurably.

I am equally thankful for the opportunity to mentor master's students, particularly Evan Fellman, learning as much from them as they did from me. The PhD Lounge was a safe space when my shoulders were sagging, where I was always able to stand straight again. Thank you, Youngseog Chung, Ian Char, Viraj Mehta, Ritesh Noothigattu, Elan Rosenfeld, and Avinava Dubey for getting me through the toughest of it with the games of Super Smash Bros or Ping Pong. I'd also like to extend a major thank you to Diane Stidle for her graceful assistance in navigating the mountains of paperwork and expectations associated with this PhD.

Our Fall of 2018 PhD Cohort will always hold a special place in my heart. There's no one else with whom I'd rather have spent countless all-nighters, battling against the relentless tide of homework, projects, and paper deadlines. A heartfelt shoutout to the dauntless spirits–Youngseog Chung, Jonathan Byrd, Ojash Neopane, Ian Char, Conor Igoe, Robin Schmucker, Ben Eysenbach, Jeremy Cohen, Tom Yan, Terrance Liu, Euxhen Hasanaj, and Théophile Gervet. Your willingness to not only burn the midnight oil in pursuit of our academic endeavors but also to rise with the sun for biking, climbing, and other exhilarating adventures embodies the essence of camaraderie and resilience. Your steadfast companionship made every challenging moment and lofty deadline not just bearable, but memorable.

Outside of academia, my life in Pennsylvania was greatly enhanced by friends and adventures that provided much-needed breaks and perspective. Special thanks to Ronnie Burdette, Foster Tucker, Kyler Tucker, Heather Tucker, Jeremy Virgin, Ethan Flannigan, Cole Browning, Victor Nardini, Mike and Emily Trombly, Brian and Emily Smith, and Jeremy Zang for dragging me on wild adventures to get my mind off of things when needed most.

During moments of uncertainty and aimlessness within my PhD journey, dirt biking emerged not just as a pastime but as a beacon, guiding me back to clarity and purpose. This adventure rekindled my bond with the tangible world and unexpectedly bridged my passion for artificial intelligence and the dynamism of motorcycle racing. I owe immense gratitude to Brian Sizemore, Chance Grillot, Youngseog Chung, and Motolani Olerinre, who were instrumental in the creation of Performance Photo - a web app that allowed me to deliver cutting-edge machine learning models for motorcycle race photo sharing, and gathering necessary data for several intriguing research challenges and questions. It became a pivotal tool that fueled my perseverance through the demanding final chapters of my PhD journey.

Ultimately, the cornerstone of my journey has been my family, whose unwavering support has been my constant. My mother's encouragement never faltered, offering tactful strength and endless comfort during my darkest moments. My father provided essential grounding and invaluable assistance through every venture. My grandparents were a source of wisdom and comfort, always ready with perspective and a warm meal. My siblings were the ever-present reminder of our collective resilience and capability. My wife has been a steadfast beacon, keeping me focused on our shared goals and the significance of our sacrifices. And my son, the purest embodiment of my motivations, constantly reminds me of life's true priorities and the reasons behind my relentless effort.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The quintessence of machine learning (ML) has traditionally been anchored in the domain of supervised multi-class classification, where models are tasked with assigning each input to one of several pre-defined categories (Shalev-Shwartz and Ben-David, 2014). This method forms the foundation of many contemporary ML systems (Russell and Norvig, 2010). From classic techniques like support vector machines (SVMs) to classify stock market trends, to the cutting-edge large language models (LLMs) that forecast the next series of words, supervised multi-class classification is the common thread on which various ML applications are based[1].

While multi-class classification has significantly shaped ML applications, there is a growing shift towards more generalized scenarios (Liu et al., 2020a; Wang et al., 2020b; Farahani et al., 2021). As these settings evolve, there is a heightened demand for methods that can handle entirely new groups of inputs or refine existing groups into more detailed subcategories. Against this backdrop, contrastive learning emerges as a particularly compelling framework (Garg et al., 2023). Unlike traditional classification, which evaluates a single input for categorization, contrastive learning examines pairs of inputs to determine their *similarity*. This approach offers a more flexible structure for understanding and organizing data.

Contrastive learning offers greater flexibility compared to traditional classification methods. Essentially, it allows any multi-class classification issue to be reinterpreted as a pair-wise comparison problem. This is done by comparing an unfamiliar input against known samples from each class and identifying the class to which the unknown most closely aligns, as illustrated in Figure 1.1. Beyond mere classification, this framework supports the optimization of various other criteria. Namely, recent advancements in large language models (LLMs) have led to the proliferation of retrieval augmented generation (RAG) systems. While the generation part of these systems is often highlighted, the retrieval component is equally critical. Effec-

[1] Although the pretraining phase of large language models is often referred to as self-supervised, this mainly reflects the self-evident labeling inherent in the training data. Ultimately, the process used to fine-tune the parameters of an LLM remains rooted in traditional supervised multi-class classification techniques.

tive retrieval is predicated on the ability to store information meaningfully, achieved through embedding models (which are often based on LLMs themselves) optimized with contrastive techniques. These models transform information into a latent space, enabling the retrieval of the most relevant data by matching the encoded query with similar items in this space during retrieval. This approach has significantly influenced a range of fields beyond natural language processing (NLP), including image retrieval, facial recognition, and the detailed examination of social media data.

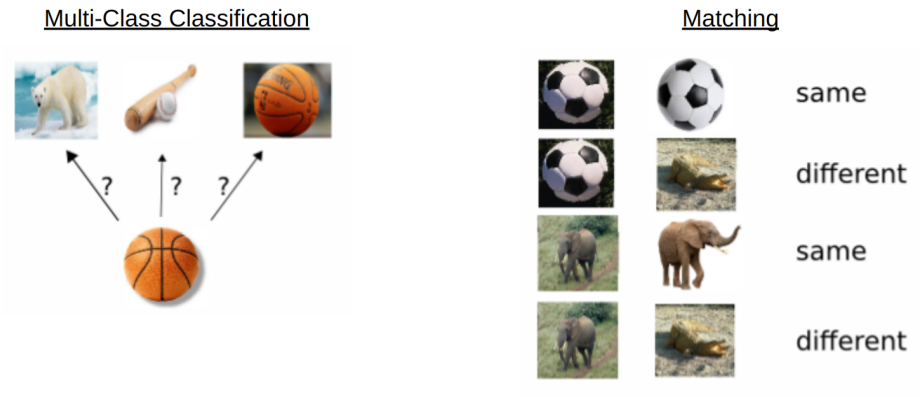


Figure 1.1: Graphical depiction of the standard classification setting versus the contrastive (or matching) setting. The standard classification setting (left) highlights a single input (the basketball) where a model must predict which of the classes it belongs. The contrastive setting (right) highlights four pairs of inputs, in each case the model predicts if the inputs are the same or different. Images from http://www.cs.toronto.edu/~rsalakhu/papers/oneshot1.pdf.

To further elucidate Figure 1.1, consider the form of the model outputs. Figure 1.2 highlights the core difference in approach: standard classification directly aims to categorize inputs by optimizing for a distinct probability distribution across labels, whereas contrastive learning focuses on structuring a latent space where the relational distances between points (representing data) convey significant, interpretable meaning. This conceptual shift underlines the contrastive approach's unique advantage in scenarios where understanding and leveraging the relationships between data points is crucial.

This thesis centers around contrastive learning, the cornerstone technique that strategically embeds inputs within a latent space to facilitate efficient and accurate retrieval. The development of models that are not only powerful but also adaptable requires a nuanced understanding of how latent spaces can be fine-tuned and optimized. However, contrastive learning is not without its challenges. One significant issue is defining what constitutes similarity: *what does it mean for two inputs to be considered "the same"?* This question is deceptively complex, as there are numerous valid interpretations, and it is not always clear which interpretation is most relevant to a given model's objectives.

The distinct features of contrastive learning bring several advantages. Firstly, it allows for a more versatile training approach; rather than solely relying on traditionally labeled data, it can

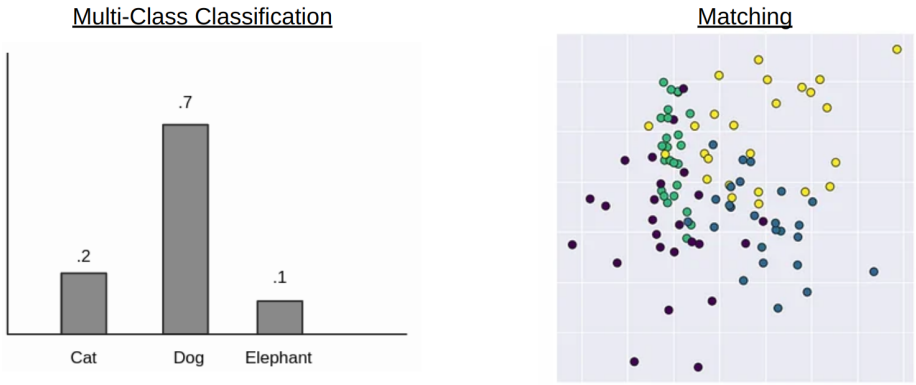


Figure 1.2: A classification model directly optimizes for a probability distribution over labels (left), whereas a contrastive model optimizes directly for a latent space (right), where the distances are used to determine a classification or other output of interest. Images from http://www.cs.toronto.edu/~rsalakhu/papers/oneshot1.pdf.

harness data presented in pairs. This flexibility enhances the learning process, especially when direct labels are scarce or incomplete. Secondly, contrastive learning is specifically designed to improve the efficiency of search and retrieval tasks. This is particularly beneficial when an input needs to be processed by a model a single time yet queried repeatedly, necessitating a profound understanding of the relationships between different instances and proving invaluable in scenarios with a vast number of classes. Lastly, this approach inherently supports zero-shot and few-shot learning capabilities. In situations where only a minimal amount of data is available for each class, or when the class labels are predetermined, the latent space created by contrastive models is generally robust enough to meaningfully incorporate and recognize new, unseen examples.

This thesis conducts a comprehensive exploration of both the advantages and limitations of contrastive models. We begin with an examination of relevant applications, such as authorship identification, where the goal is to determine the author of a text using only the written content, and person search, which involves identifying and matching images of an individual across time and space. A significant portion of this investigation culminates in the practical implementation of a photo-searching system designed for off-road motorcycle racing. This system tests the limits of existing person search models under challenging conditions, as illustrated by Figure 1.3, which displays four images of the same racer captured only minutes apart. The real-world challenges identified through these applications have informed the development of this thesis, leading to the proposal of new methodologies aimed at enhancing the efficacy of contrastive models. These methodologies focus on optimizing sample weightings and addressing the complexities associated with learning from weakly labeled data.

Figure 1.3: Four images of a single racer, taken only minutes apart during a muddy race.

1.1 Thesis Statement and Overview

This thesis delves into the dynamics and challenges inherent in training contrastive models. At its core, the investigation is guided by the following central thesis:

> *To build machine learning systems that are not only reliable but also practically beneficial for real-world applications, it is imperative that we:*
>
> 1. *Thoroughly understand and differentiate the characteristics and implications of contrastive versus non-contrastive learning paradigms.*
> 2. *Innovate and refine methods and algorithms to synergistically integrate the strengths of both paradigms.*

The thesis is organized into two main parts, each exploring a facet of the thesis statement.

Part I: Contrastive Learning in the Real World

The first part of this thesis is dedicated to exploring the practical applications and real-world implications of contrastive learning. Here, we delve into three primary areas: authorship attribution (AA), authorship verification (AV), and person re-identification (ReID). Each of these domains presents unique challenges and opportunities for the application of contrastive learning methodologies, and together they illustrate the broad utility and adaptability of these techniques in different contexts.

Chapter 2 is focused on the use of Siamese BERT models for authorship verification. This chapter introduces the concept of using Transformer-based models, specifically BERT, for comparing texts to determine authorship. We explore the nuances of feature extraction and the importance of capturing stylistic elements unique to individual authors. The chapter provides a foundational understanding of how contrastive learning can enhance the capabilities of NLP systems in identifying and distinguishing between authors' styles.

Chapter 3 expands on the theme of authorship analysis by presenting a comprehensive framework for standardizing and benchmarking authorship attribution and verification. This involves an empirical evaluation and comparative analysis of different methodologies within this field. The chapter makes the differences between contrastive and classification models tangible and aims to establish a common ground for future research in authorship analysis, enabling more consistent and transparent comparisons between different approaches.

Chapter 4 shifts the focus to a visually oriented application: text spotting in off-road motorcycle racing. This chapter addresses the challenges of applying machine learning to dynamic and unpredictable environments, such as those encountered in off-road racing. By introducing a new dataset tailored to these conditions, we illustrate the difficulties of conventional OCR and ReID models when faced with extreme variability and propose methodologies for enhancing their robustness. While this is not a direct application of contrastive learning, the signal generated from the numbers present on off-road racers is later used to enhance the contrastive power of off-road racer re-identification models.

Chapter 5 continues the exploration of machine learning in sports analytics, specifically through the lens of person re-identification in off-road motorcycle racing. This chapter discusses the creation of MUDD, a new dataset designed for ReID tasks under extreme conditions. The focus is on the unique challenges posed by mud, dust, and other environmental factors, and how contrastive learning can be leveraged to improve model performance in such settings.

Through these chapters, Part 1 of the thesis demonstrates the versatility and effectiveness of contrastive learning across a variety of domains and introduces new datasets for improving contrastive learning in difficult scenarios. By examining its application in both textual and visual contexts, we aim to provide a comprehensive overview of how contrastive learning methods can be adapted and applied to solve specific real-world challenges. This part of the thesis underscores the importance of understanding the characteristics of contrastive versus non-contrastive systems in building reliable and practically useful machine learning systems suitable for real-world deployment.

Part II: Learning to Improve Contrastive Learning

The second part of this thesis aims to advance our understanding and methodologies within the realm of contrastive learning itself. It is dedicated to developing new techniques and algorithms that enhance the effectiveness and efficiency of contrastive learning models. Part I revealed the importance of mining functions in contrastive learning. Accordingly, half of this part is focused on addressing the challenges associated with mining functions and risk optimization

in the context of machine learning, thereby capturing the best attributes of both contrastive and non-contrastive learning systems. Another major shortcoming identified in Part I was the lack of contrastive-based methodologies for training among very high levels of noise. The second half of this part introduces new methods for alleviating this pain point.

Chapter 6 introduces a novel approach to optimizing risk functions in machine learning, particularly focusing on the concept of Risk-Adjusted Mini-batches (RAM). This chapter explores how meta-learning can be employed to learn specialized and interpretable reweightings over mini-batches, thereby minimizing any differentiable risk function. By proposing a new optimization procedure, this chapter addresses the challenges faced in typical mini-batch learning, especially when dealing with complex risk functions. This approach broadens our understanding of risk management in machine learning applications but also provides a major benefit for contrastive learning – the removal of hand-engineered mining functions in favor of those learned.

Chapter 7 introduces Contrastive Multiple Instance Learning (CMIL). This chapter tackles the problem of learning from weakly labeled data, a common issue in real-world machine-learning scenarios. By extending multiple-instance learning with contrastive learning principles, this methodology introduces a new way of handling ambiguously labeled groups of photos. This is particularly relevant in scenarios like person re-identification, where traditional models struggle with sparse and noisy labels. The chapter provides a detailed analysis of CMIL's effectiveness and its potential to revolutionize the way we approach learning in less-than-ideal labeling conditions.

Throughout Part 2, the thesis not only addresses the theoretical underpinnings and practical applications of these advanced methodologies but also demonstrates how these innovations contribute to the broader goal of improving the reliability and utility of contrastive learning systems. By exploring new frontiers in risk optimization and weakly supervised learning, this part of the thesis contributes significantly to the field, offering novel insights and tools that can be applied across a spectrum of machine learning challenges.

1.2 How to Read This Thesis

The structure of this thesis is crafted to facilitate a comprehensive understanding of the subject matter, with the content thoughtfully divided into two distinct parts. While each section and chapter has been designed to stand alone—offering self-contained insights and value—certain chapters naturally complement each other and are thus best consumed in conjunction.

Specifically, Chapter 3 builds directly upon the foundational concepts introduced in Chapter 2, creating a synergistic duo that thoroughly explores the realm of authorship attribution and verification. This combination serves as an ideal preliminary exploration that sets the stage for the advanced discussions that follow, particularly in Chapter 6. Together, these chapters weave

a narrative that underscores the pivotal role of mining functions, transitioning smoothly from practical, real-world applications to the more abstract, theoretical underpinnings and the exploration of innovative mining function methodologies.

Conversely, Chapters 4 and 5 are intricately linked, laying the foundational groundwork for the examination of machine learning applications in the challenging environment of off-road motorcycle racing. These chapters, when read in sequence, provide a deep dive into the unique challenges and innovative solutions associated with text spotting and person re-identification under extreme conditions. Upon these insights, Chapter 7 introduces strategies for addressing weakly labeled data of the exact nature as presented in the off-road racing datasets.

1.3 Bibliographic Notes

Most of the work in this thesis is based on the following papers:

Part I, Chapter 2 is based on:

- Tyo, Jacob, Bhuwan Dhingra, and Zachary C. Lipton. "Siamese Bert for Authorship Verification." CLEF (Working Notes). 2021.

Part I, Chapter 3 is based on:

- Tyo, Jacob, Bhuwan Dhingra, and Zachary C. Lipton. "Valla: Standardizing and Benchmarking Authorship Attribution and Verification Through Empirical Evaluation and Comparative Analysis." Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.

Part I, Chapter 4 is based on:

- Tyo, Jacob, et al. "Reading Between the Mud: A Challenging Motorcycle Racer Number Dataset." arXiv preprint arXiv:2311.09256 (2023). (Under Submission)

Part I, Chapter 5 is based on:

- Tyo, Jacob, et al. "MUDD: A New Re-Identification Dataset with Efficient Annotation for Off-Road Racers in Extreme Conditions." arXiv preprint arXiv:2311.08488 (2023). (Under Submission)

Part II, Chapter 6 is based on:

- Tyo, Jacob, and Zachary C. Lipton. "Meta-Learning Mini-Batch Risk Functionals." arXiv preprint arXiv:2301.11724 (2023). (Under Submission)

Part II, Chapter 7 is based on:

- Tyo, Jacob, and Zachary C. Lipton. "Contrastive Multiple Instance Learning for Weakly Supervised Person ReID." arXiv preprint arXiv:2402.07685 (2024). (Under Submission)

Part I

Contrastive Learning in the Real World

Chapter 2

Siamese BERT for Authorship Verification

The PAN 2021 authorship verification (AV) challenge focuses on determining if two texts are written by the same author or not, specifically when faced with new, unseen, authors. In our approach, we construct a Siamese network initialized with pretrained BERT encoders, employing a learning objective that incentives the model to map texts written by the same author to nearby embeddings while mapping texts written by different authors to comparatively distant embeddings. Additionally, inspired by related work in computer vision, we attempt to incorporate triplet losses but are unable to realize any benefit. Our method results in a slight performance gain of 0.9% overall score over the baseline and an increase of 8% in F1 score.

2.1 Introduction

Authorship verification (AV) is the task of determining if two texts were written by the same person or not. While traditionally, this feat has required the expertise of forensic linguists, recent advances in both natural language processing (NLP) and related matching tasks in computer vision, offer several paths for improving automated methods. The traditional machine learning approach to this problem consists of two steps: feature extraction and model fitting. Feature extraction can include the count of specific words/sub-words/punctuation, misspellings, part-of-speech tags, etc. More recent methods have paired these hand-engineered features with modern feature extraction methods such as n-grams Ruder et al., 2016, pretrained word embeddings Boenninghoff et al., 2020, and pretrained sentence structure embeddings Jafariakinabad and Hua, 2020. The models leveraged for this task have ranged from latent Dirichlet allocation Savoy, 2013 and support-vector machines Campo-Rodríguez et al., 2018 to convolutional Ruder et al., 2016; Shrestha et al., 2017 and recurrent neural networks Bagnall, 2015; Jafariakinabad et al., 2019. However, prior work in AV has not yet made extensive use of trans-

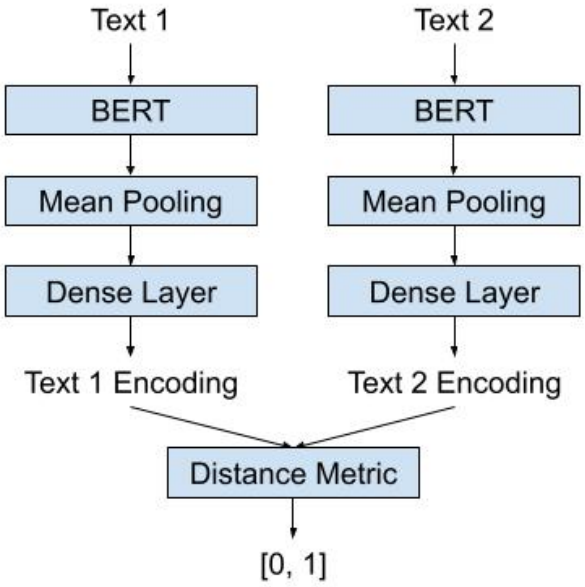


Figure 2.1: The Siamese BERT for Authorship Verification (SAV) model structure in the givand data flow.

former architectures or pretrained language models.[1]

In this work, we apply the pretrained BERT model in a Siamese configuration for the task of AV Bevendorff et al., 2021. We make use of WordPiece Wu et al., 2016 for tokenization and do not use engineered features. We set out to determine how well modern methods perform on AV, and the feasibility of removing hand-engineered features in favor of deeper models and prior knowledge in the form of pretraining. Furthermore, triplet loss has provided benefits in image processing Hermans et al., 2017, but has not yet been leveraged for AV. We experiment with triplet loss (leveraging multiple sampling strategies), contrastive loss, and a modified version of contrastive loss that has proved beneficial in a previous AV study Boenninghoff et al., 2019. The dataset for this task was obtained from fanfiction.net, where each datapoint consists of pairs of text from two different fanfics (an amateur fictional writing based on an existing work of fiction) Bevendorff et al., 2021. More on this dataset in Section 2.2.1.

2.2 Siamese BERT for Authorship Verification

We introduce Siamese BERT for Authorship Verification (SAV)[2]. Our method uses a pretrained BERT model in a Siamese setup as shown in Figure 2.1 and originally introduced by Reimers and Gurevych, 2019a. In the AV task, we are given two input texts x_1 and x_2 and the expected output is a score in the interval $[0, 1]$ indicating the likelihood with which they belong to the same author. The maximum input size for the BERT model is 512 tokens, therefore we truncate each text to the first 512 tokens. Separately, for both input texts, they are passed through the BERT model resulting in an output of size $n \times 768$ (where n is the number of tokens in the input and 768 is the dimension of the BERT output for each token). All n representations are then averaged into a 768 dimensional vector (the mean pooling layer), and then passed through a fully connected layer to generate the final text embedding (256 dimensional). This gives the final output representation u and v of input texts x_1 and x_2, respectively. These representations are then compared using a distance metric, which is then used for loss calculation and model optimization.

During inference, the same procedure is followed. The only difference is that after the distance between embeddings u and v are calculated, it is compared to a threshold. If the corresponding distance is smaller than the threshold, the texts are predicted to have been written by the same author and vice versa. In Section 2.4 we discuss more detail on finding the thresholds, as well as an alternative approach to truncating each input to 512 tokens.

2.2.1 Training

Data Preprocessing

The PAN 2021 AV challenge provided two datasets, both obtained from fanfiction.net. Each datapoint consists of a pair of texts from two different fanfics, as well as a tag representing which fandom (the particular fictional series) each text is from. We leverage only the large dataset in this work, which contains $275,565$ text pairs. Roughly 54% of these pairs were written by the same author (i.e. a same-author pair). Approximately 8% of the pairs were texts from the same fandom, but none of the same-author pairs contained texts from the same fandom. In total, the texts were pulled from $1,600$ fandoms and over $278,000$ authors.

Instead of using these predefined pairs for training, we elected to split all pairs and store all texts individually. However, we don't want to change the data distribution for the test set. Therefore, we sample 10% of the pairs randomly to form the test set. We then ensure that all

[1]We note that several works in Authorship Attribution (classifying texts into a fixed list of potential authors) do leverage pretrained language models, including Barlas and Stamatatos, 2020; Fabien et al., 2020; Fourkioti et al., 2019.

[2]All code for this model can be found here: https://github.com/JacobTyo/PAN21_SAV

authors found in this test set have no texts in the training set. If so, the text pair is moved to the test set. We form a secondary test set by splitting all of the test pairs, and then recombining them randomly (based on author, using the same procedure as is used during training). We will refer to this set as the modified test set, as it has the same data distribution as the training set we have created but not as the original data. During training, we randomly sample text pairs (at roughly 50% same-author 50% different author pair rates). Although this changes the data distribution, it allows us to leverage a much larger set of text pairs (≈76 billion possible pairs vs ≈275 thousand).

Loss Functions

Any Siamese model can be trained using a wide range of loss functions. In this work, we explored training our model with the contrastive, modified contrastive, and triplet loss functions.

The contrastive loss is

$$\mathcal{L}_{\text{c}}(u, v, y) = \frac{1}{2}\Big(y\ \text{d}(u, v)^2 + (1 - y)\max\{(m - \text{d}(u, v))^2, 0\}\Big), \tag{2.1}$$

where u and v are text embeddings, $y \in \{0, 1\}$ is the label (1 if u and v were written by the same author, 0 otherwise), d is the distance metric, and m is a margin (no loss is incurred for a different-author pair if their representations are further apart than m).

The modified contrastive loss, originally introduced by Boenninghoff et al., 2019, is

$$\mathcal{L}_{\text{mc}}(u, v, y) = \frac{1}{2}\Big(y\max\{(\text{d}(u, v) - m_s)^2, 0\} + (1 - y)\max\{(m_d - \text{d}(u, v))^2, 0\}\Big). \tag{2.2}$$

The modification from the aforementioned contrastive loss is that there are now two margins. m_s refers to a margin for same-author pairs. If the distance between the embeddings of a same-author pair is smaller than m_s, then no loss is incurred. In normal contrastive loss, loss is incurred unless the pair of texts evaluate to identical representations. This modified contrastive loss allows for some variation among the texts of a single author, which should help account for differences in a single author's text such as topic differences, and therefore make the resulting model more robust to non-stylistic difference among authors. The second margin m_d refers to a margin for different-author pairs, and performs the same function as m in the original contrastive loss.

The triplet loss function is

$$\mathcal{L}_{\text{t}}(a, p, n) = \max\{d(a, p) - d(a, n) + m, 0\}, \tag{2.3}$$

where a represents the embedding of an *anchor* text, p represents the embedding of a different text than a but from the same author (*positive* pair), and n represents the embedding of a text from an author different than that of a (*negative* pair). m is the margin to separate the positive

and negative pairs by (i.e. the negative sample should be further from the anchor than the positive sample by at least m). Note that the triplet loss does not explicitly push same-author pairs together, but instead only forces different-author pairs to be farther apart than same-author pairs. With the contrastive and modified contrastive loss functions, we sample pairs of texts randomly. With triplet loss, it is common to use different sampling techniques. Hermans et al. (2017) describe an efficient way of performing hard negative mining. Given a random batch of samples, the loss is computed (according to the triplet loss function) for all possible, valid triplets. Then the hardest positive and the hardest negative (i.e. the positive that is furthest from the anchor and the negative that is closest to the anchor) are selected, and the loss with respect to these samples is used for updating.

Distance Metrics

We test with the cosine ($\mathrm{d}_{\cos}$) and Euclidean (d_{euc}) distance measures:

$$\mathrm{d}_{\cos}(u, v) = 1 - \frac{u \cdot v}{||u||\ ||v||} \tag{2.4}$$

$$\mathrm{d}_{\text{euc}}(u, v) = ||u - v||_2 \tag{2.5}$$

Resources

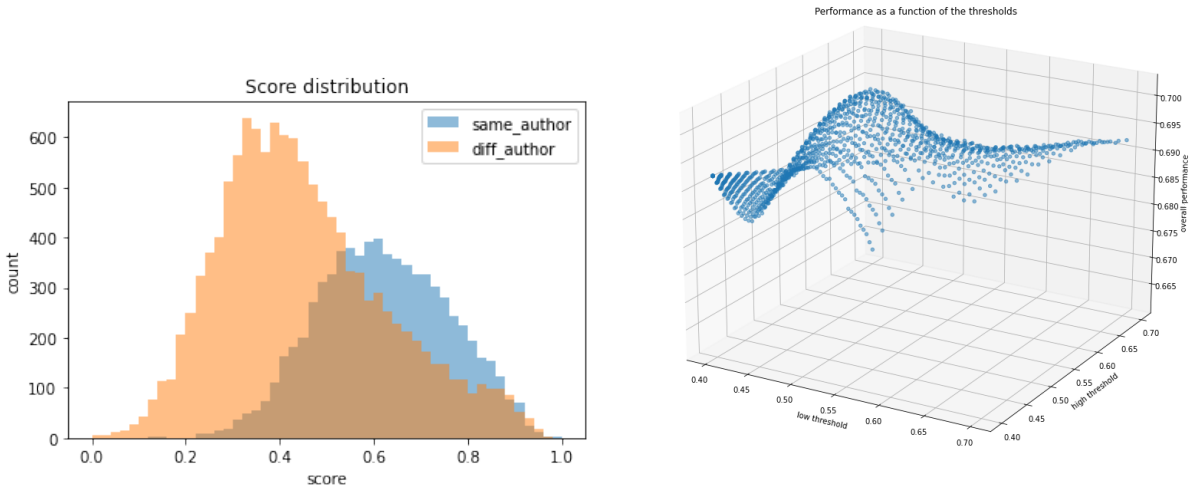
The final model was trained for 3 days on 8 Tesla v100's. This allowed for 16 samples per GPU, for a total batch size of 128. We used the standard learning rate for the hugging face transformer pretrained models (5×10^{-5}) and anneal it over 4 epochs.

2.3 Evaluation

We use the evaluation metrics described in Kestemont et al., 2020, as well as the baseline model provided as part of the AV task[3] Kestemont et al., 2021b:

- AUC: the conventional area-under-the-curve of the precision-recall curve
- F1-score: the harmonic mean of the precision and recall Pedregosa et al., 2011
- c@1: a variant of the conventional F1-score, which rewards systems that leave difficult problems unanswered (i.e. scores of exactly 0.5) Peñas and Rodrigo, 2011

[3]As described in Kestemont et al., 2020, the provided baseline is a simple method that calculates the cosine similarities between TF-IDF-normalized, bag-of-character-tetragrams representations of the texts in a pair. The resulting scores are then shifted using a simple grid search, to arrive at an optimal performance on the calibration data.



(a) (b)

Figure 2.2: (a) The score distribution of the final model on the test set. (b) The overall performance (z-axis) with respect to the lower and upper thresholds (x-axis and y-axis respectively).

- F_0.5u: a newly proposed measure that puts more emphasis on deciding same-author cases correctly Bevendorff et al., 2019
- overall: the simple average of all previous metrics

For hyperparameter selection, we predefined 17 models that differ in terms of their loss function, distance metric, and margin(s). Each of these models is trained for 3 days on a single Tesla V100 GPU. Table 2.2 details the performance of each of these models with respect to the modified testing set. The highest performing model with respect to the overall score is one that leverages the modified contrastive loss along with the Euclidean distance metric and an upper and lower margin of 5 and 0.25 respectively. We choose this hyperparameter combination for our final model, which was then evaluated on a hidden test set via the TIRA environment Potthast et al., 2019. Table 2.1 shows the performance of our model on this hidden test set.

Table 2.1: The performance of our final model on the hidden test set, evaluated on the TIRA environment for the PAN21 competition Kestemont et al., 2021b

Model	AUC	F1	c@1	F_0.5u	Brier	Overall
Final Model	0.8275	0.7911	0.7594	0.7257	0.8123	0.7832

Table 2.2: The performance of all models during the hyperparameter search. All models were trained for 3 days on a single Tesla V100 GPU, and were evaluated on the modified test set. The "Final Model" entry details the performance of the best performing model on the modified test set after completing the final training phase (3 days on 8 Tesla V100's - i.e. 8x larger batch size for the same number of training iterations).

Loss Function	Distance Metric	Upper Margin	Lower Margin	AUC	F1	c@1	F_0.5u	Overall
Triplet	Euclidean	1000	-	0.552	0.206	0.522	0.341	0.405
		100	-	0.561	0.361	0.537	0.468	0.482
		10	-	0.529	0.457	0.519	0.495	0.500
Contrastive	Cosine	0.01	-	0.878	0.692	0.556	0.584	0.678
		0.1	-	**0.904**	0.696	0.563	0.588	0.688
		0.5	-	0.874	0.795	0.773	0.751	0.798
		0.9	-	0.826	0.646	0.712	0.749	0.733
Contrastive	Euclidean	0.1	-	0.839	0.765	0.738	0.72	0.766
		1	-	0.805	0.737	0.722	0.712	0.744
		10	-	0.782	0.663	0.699	0.715	0.715
Modified Contrastive	Euclidean	0.5	2.5	0.813	0.743	0.723	0.712	0.748
		0.25	5	0.871	**0.809**	0.776	0.757	**0.803**
		0.5	5	0.789	0.722	0.705	0.698	0.729
		0.25	1	0.881	0.789	0.754	0.688	0.778
		0.1	1	0.858	0.778	0.756	0.713	0.776
		0.75	1	0.815	0.627	0.704	0.698	0.711
Modified Contrastive	Cosine	0.1	0.5	0.811	0.719	0.729	0.735	0.749
Baseline	-	-	-	0.831	0.769	**0.78**	**0.764**	0.786
Final Model	-	-	-	0.892	0.811	0.796	0.777	0.819

2.4 Analysis of the final model

The final model was trained for 3 days on 8 Tesla V100 GPU's, equating to roughly 4 epochs. All analysis and results are with respect to the test set, not the modified test set as in previous sections. We first examine the score distribution, shown in Figure 2.2a. The same author pairs are represented by the blue histogram, and different author pairs are represented by the yellow. There is still significant overlap in the scores of the two different groups.

We optimize our thresholds on the test set via grid search, which is visualized in Figure 2.2b. The z-axis represents the overall performance of the final model as the thresholds are varied (x and y axes). The optimal thresholds are 0.470 and 0.553 respectively, giving an overall performance on the test set of 0.701. The other performance metrics, along with the performance of the baseline on test set is shown in Table 2.3

2.4.1 Adding more context

Our model leverages only the first 512 tokens of each text (the maximum tokens that will fit in a single pass through the BERT model). Here, using the final model, we investigate chunking each text longer than 512 tokens into sets, and then combining the final representation of each chunk before passing the text representation to the distance metric. We combine the representations of each individual chunk via averaging, resulting in a fixed-length vector that encodes information from the entire input text regardless of length. The score distribution for the final model using chunking is shown in Figure 2.3. We compare these two score distributions by looking at the percentage of overlap, and find that both the chunking and non-chunking procedures produce roughly 55% overlap. Furthermore, this chunking behavior results in slightly worse overall performance on the test set, 0.701 vs 0.715 of the final model without chunking. Lastly, chunking is computationally expensive. During inference, on an 8-core CPU, the final model takes about 1.5 seconds to process an input pair. On the same machine, the chunking model takes about 9.8 seconds to process that same input pair. Because of this large cost increase along with the indistinguishable performance, we use only the non-chunking model.

2.5 Conclusion

In this work, we construct a Siamese network initialized with pretrained BERT encoders, employing a learning objective that incentives the model to map texts written by the same author to nearby embeddings while mapping texts written by different authors to comparatively distant embeddings. Our method results in a slight performance gain over a baseline of 0.9% overall score, and an increase of 8% in F1 score. We explore the effectiveness of different loss functions, distance metrics, and margins and our results indicate the need for either hand engineered features or more training time and data. This work represents the first steps in understanding the ability of modern language models and tokenizers to perform authorship verification, without any of the common hand engineered features. Some interesting future work includes broadening the training data (incorporating many AV datasets during training) and lengthening the training time, further investigating sampling strategies (we expect approaches such as hard-negative mining to provide improvements vs random sampling), and explore different methods of embedding text longer than the input size of the BERT model.

Table 2.3: The performance for the baseline and final model on the test set.

Model	AUC	F1	c@1	F_0.5u	Overall
Baseline	0.779	0.659	0.759	0.628	0.706
Final Model	0.780	0.739	0.731	0.611	0.715

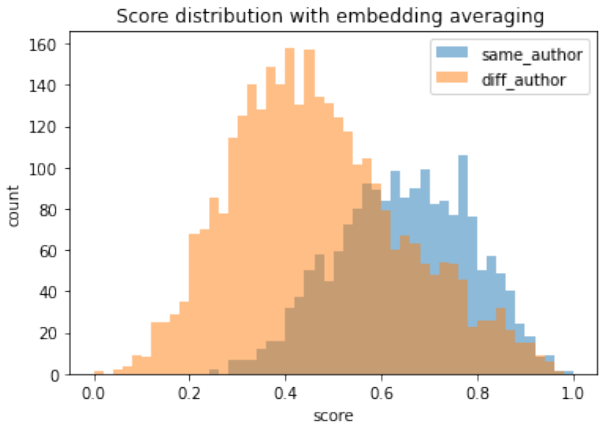


Figure 2.3: The score distribution of the final model on the test set when using chunking.

Chapter 3

VALLA: Authorship Identification Benchmark

Despite decades of research on authorship attribution (AA) and authorship verification (AV), inconsistent dataset splits/filtering and mismatched evaluation methods make it difficult to assess the state of the art. In this paper, we present a survey of the fields, resolve points of confusion, introduce VALLA that standardizes and benchmarks AA/AV datasets and metrics, provide a large-scale empirical evaluation, and provide apples-to-apples comparisons between existing methods. We evaluate eight promising methods on fifteen datasets (including distribution shifted challenge sets) and introduce a new dataset based on texts archived by Project Gutenberg. Surprisingly, we find that a traditional Ngram-based model performs best on 5 (of 7) AA tasks, achieving an average macro-accuracy of 76.50% (compared to 66.71% for a BERT-based model). However, on the two AA datasets with the greatest number of words per author, as well as on the AV datasets, BERT-based models perform best. While AV methods are easily applied to AA, they are seldom included as baselines in AA papers. We show that through the application of hard-negative mining, AV methods are competitive alternatives to AA methods. VALLA and all experiment code can be found here: https://github.com/JacobTyo/Valla

3.1 Introduction

The statistical analysis of variations in literary style between one writer or genre and another, commonly known as *stylometry*, dates back as far as 500 AD. Computer-assisted stylometry first emerged in the early 1960s, when Mosteller and Wallace (1963) explored the foundations of computer-assisted authorship analysis. Today automated tools for authorship analysis are common, finding practical use in the justice system to analyze evidence (Koppel et al., 2008), among social media companies to detect compromised accounts (Barbon et al., 2017), to link online accounts belonging one individual (Sinnott and Wang, 2021), and to detect plagiarism (Stamatatos and Koppel, 2011).

In the modern Natural Language Processing (NLP) literature, two problem formulations dominate the study of methods for determining the authorship of anonymous or disputed texts: Authorship Attribution (AA) and Authorship Verification (AV). In AA, the learner is given representative texts for a canonical set of authors in advance, and expected to attribute a new previously unseen text of unknown authorship to one of these a priori known authors. In AV, the learner faces a more general problem: given two texts, predict whether or not they were written by the same author.

While both problems have received considerable attention (Murauer and Specht, 2021; Altakrori et al., 2021; Kestemont et al., 2021a), the state of the art is difficult to assess owing to inconsistencies in the datasets, splits, performance metrics, and variations in the framing of domain shift across studies. For example, a recent survey paper (Neal et al., 2017) indicates that the state-of-the-art method is based on the Prediction by Partial Matching (PPM) text compression scheme and the cross-entropy of each text with respect to the PPM categories. By contrast, the PAN-2021 competition (Kestemont et al., 2021a) indicates that the state of the art is a hierarchical bi-directional LSTM with learned-CNN text encodings. Recent work (Fabien et al., 2020) concludes that the transformer-based language model BERT is the highest-performing AA method. A recent analysis paper (Altakrori et al., 2021) argue that the traditional approach of character n-grams and masking remains the best methodology to this day. Each of these sources compares methods against different baselines, on different datasets (sometimes on just a single small dataset), and with different problem variations (such as cross-topic, cross-genre, etc.).

In this paper, we start by sorting out this fragmented prior work through a brief survey of the literature. Then, to present a unified evaluation, we introduce Valla. Valla provides standardized versions of all the common AA and AV datasets with uniform evaluation metrics and standardized domain-shifted test sets, and implementations of all methods used in this paper. Additionally, we introduce a new large-scale dataset based on public domain books sourced from Project Gutenberg for both tasks. Then using this benchmark, we present an extensive evaluation of eight common AA and AV methods on their respective datasets with and without domain shift. We also make comparisons *between* AA and AV methods where applicable.

Recent work indicates that traditional methods still outperform pretrained language models (i.e. BERT) (Kestemont et al., 2021a; Altakrori et al., 2021; Murauer and Specht, 2021; Tyo et al., 2021; Peng et al., 2021; Futrzynski, 2021), but we show that this narrative only appears to apply to datasets with a limited number of words per class. Furthermore, BERT-based models achieve new state-of-the-art macro-accuracy on the IMDb62 (98.80%) and Blogs50 (74.95%) datasets and set the benchmark on our newly introduced Gutenberg dataset.

The applicability of AV methods to AA problems is frequently mentioned, yet these methods are not placed in competition. We provide this comparison and find that AA methods to outperform AV methods on AA problems, *but only until* hard-negative mining is used during AV training. Initially, AA outperform AV methods by 15% macro-accuracy, but hard negative min-

Dataset	Text Type	Typical Setting	iid	$\times_t$	$\times_g$	$\times_a$	D	A	W	D/A	W/D	imb
CCAT50	News	AA	✓	—	—	—	5k	50	2.5M	100	506	0
CMCC	Various	AA	✓	✓	✓	—	756	21	454k	36	601	0
Guardian	Opinion	AA	✓	✓	✓	—	444	13	467k	34	1052	6.7
IMDb62	Reviews	AA	✓	—	—	—	62k	62	21.6M	1000	349	2.6
Blogs50	Blogs	AA	✓	—	—	—	66k	50	8.1M	1324	122	553
BlogsAll	Blogs	AV	✓	—	—	—	520k	14k	121.6M	37	233	90
PAN20 & 21	Various	AV	✓	✓	—	✓	443k	278k	1.7B	1.6	3922	2.3
Amazon	Reviews	AV	✓	✓	—	—	1.46M	146k	91.9M	10	63	0
Gutenberg	Books	AA	✓	—	—	✓	29k	4.5k	1.9B	6	66350	10.5

Table 3.1: An overview of datasets used for Authorship Attribution (AA) and Authorship Verification (AV). iid is an i.i.d. split, $\times_t$ is a cross-topic split, $\times_g$ is a cross-genre split, $\times_a$ is an unknown author split, D is the number of documents, A is the number of authors, W is the number of words, W/D is the average length of documents, D/A is the average number of documents per author, W/D is the average number of words per document, and imb is the imbalance of the dataset measured by the standard deviation of the number of documents per author. ✓ indicates necessary data is available to create a standardized split, whereas — indicates it isn't.

ing improves the performance of AV models in the AA setting, increasing the macro-accuracy of BERT_V (a verification formulation of the BERT model) to 72.42% on the tested dataset, making it a competitive alternative. In summary, we contribute the following:

- A survey of AA and AV.
- A benchmark that standardizes AA and AV datasets and method implementations
- State-of-the-art accuracy on the IMDb62 (98.80%) and Blogs50 (74.95%) datasets.
- A new dataset with long average text length.
- An evaluation of eight high-performing AA and AV methods on fifteen datasets
- Evidence of the importance of hard-negative mining for authorship applications.

3.2 Brief Survey of the Literature

Neal et al. (2017) provide an overview of AA dataset characteristics and traditional AA methods. The authors enumerate the wide array of textual features used for AA and provide an evaluation of these techniques on a single, small dataset. They conclude that the prediction using partial matching (PPM) method is the state of the art. Bouanani and Kassou (2014) provide a similar survey focusing on the enumeration of AA hand-engineered features. Stamatatos (2009) discuss traditional AA methods from an instance-based (one text vs another) vs a profile-based (one text vs all authors) methodology, and include a computational requirement analysis.

Among notable surveys, Mekala et al. (2018) compare the benefits of the different traditional textual features; Argamon (2018) detail the problems with applying many traditional AA methods in real-world scenarios; Alhijawi et al. (2018) provide a meta-analysis of the field; and Ma et al. (2020) point out the lack of advances from using transformer-based language models in AA. Critically, all of these prior surveys exclude recent advances due to deep learning, such as recurrent neural networks, transformers, word embeddings, and byte-pair encoding. In this section, we briefly cover more traditional techniques, and then discuss recent deep-learning-based approaches.

So far, we have outlined the work on AA surveys, but there are none to be found that focus on AV. The PAN competition overview Kestemont et al., 2021a is close, but limited to what appears in competition. Also of note, each year's competition focuses on a single dataset that changes every year.

3.2.1 Datasets

Murauer and Specht (2021) worked towards a benchmark for AA. They do not discuss AV or the domain shift present in many popular datasets. The test sets often contain novel topics (cross-topic - $\times_t$), genres (cross-genre - $\times_g$), or authors (unique authors - $\times_a$). Table 3.1 shows the statistical variability between the different datasets. The number of authors, documents, and words in a corpus is influential, but looking more closely at the number of documents per author (D/A) and the number of words per document (W/D) gives a better idea of how hard a corpus is. The larger the number of authors and the less text there is to work with, the harder the problem. Lastly, we measure the imbalance (imb) of datasets based on the standard deviation of the number of documents per author. The CCAT50 Lewis et al., 2004, CMCC Goldstein et al., 2008, Guardian Stamatatos, 2013, IMDb62 Seroussi et al., 2014, and PAN20 & PAN21 Kestemont et al., 2021a are used as they are in prior work, but with the distinction that we publish our train/validation/test splits to ensure comparability with future work.

Although the Blogs50 dataset (Schler et al., 2006) is common (BlogsALL in Table 3.1), the statistics we present are different than those originally published. This discrepancy is due to a large number of exact duplicates ($\sim$160,000) which we have removed. The most common form of this dataset is Blogs10 and Blogs50 (the texts only from the "top" 10 and 50 authors respectively). This is problematic because it isn't clear how these "top" authors are selected: the number of documents (Fabien et al., 2020; Patchala and Bhatnagar, 2018), the number of words, with minimum text length (Koppel et al., 2011), with spam (or other) filtering Yang and Chow, 2014; Halvani et al., 2017, or as in most cases, not specified (Jafariakinabad and Hua, 2022; Yang et al., 2018b; Zhang et al., 2018; Ruder et al., 2016). In our framework, we release standard splits and cleaning for this dataset.

Finally, we introduce the Gutenberg authorship dataset, as a new large-sclase authorship corpus with very long texts (each texts is about 17 times longer, on average, than the next

longest corpus). While some prior work has used Project Gutenberg[1] as a dataset source (public domain books), they all use small subsets (Arun et al. (2009) use 10 authors, Gerlach and Font-Clos (2020) use the 20 most prolific authors, Menon and Choi (2011) use 14 authors, Rhodes (2015) use 6 authors, Khmelev and Tweedie (2001) get a 380 text subset, etc.). Here we have collected all single-author English texts from Project Gutenberg resulting in almost 2 billion words and a very long average document length.

3.2.2 Metrics

One of the difficulties in comparing prior work is the use of different performance metrics. Some examples are accuracy (Altakrori et al., 2021; Stamatatos, 2018; Jafariakinabad and Hua, 2022; Fabien et al., 2020; Saedi and Dras, 2021; Zhang et al., 2018; Barlas and Stamatatos, 2020), F1 (Murauer and Specht, 2021), C@1 Bagnall, 2015, recall Lagutina, 2021, precision Lagutina, 2021, macro-accuracy Bischoff et al., 2020, AUC Bagnall, 2015; Pratanwanich and Lio, 2014, R@8 (Rivera-Soto et al., 2021), and the unweighted average of F1, F0.5u, C@1, and AUC (Manolache et al., 2021; Kestemont et al., 2021a; Tyo et al., 2021; Futrzynski, 2021; Peng et al., 2021; Bönninghoff et al., 2021; Boenninghoff et al., 2020; Embarcadero-Ruiz et al., 2022; Weerasinghe et al., 2021).

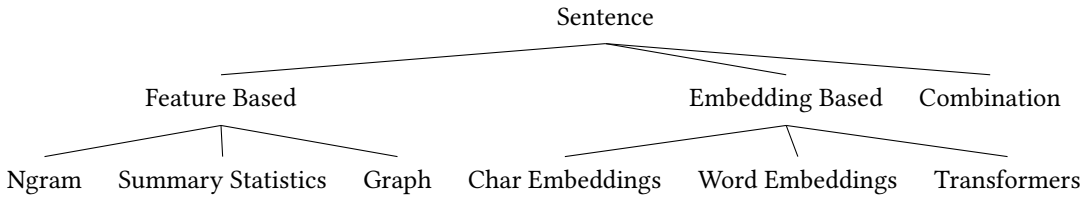


Figure 3.1: Hierarchy of feature extraction methods

In AA and AV, we want to understand the discriminative power of each model, while avoiding metrics that are influenced too much by performance on a small subset of prolific authors. Thus, we adopt *macro-averaged accuracy* for AA (referred to as macro-accuracy), and *AUC* for AV.

3.2.3 Methods

Figure 3.1 depicts our categorization.

[1]https://www.gutenberg.org

Feature Based

Ngram The most commonly seen input representation (feature) used in AA and AV problems are of N-grams. N-grams provide a fast and simple vectorization method for text that ignores order, based on a given vocabulary of tokens. Granados et al. (2011) introduced *text distortion*, which substitutes out-of-vocabulary items for a "*". Stamatatos (2018) and Bischoff et al. (2020) further test these distortion methods and more complex domain-adversarial methods, showing that the simpler distortion methods are most effective.

The Ngram-based *unmasking* method (Koppel and Schler, 2004), is based on the idea that the style of texts from the same author differs only in a few features. At its core, this method iteratively trains classifiers to predict if two texts are from the same author, but with a decreasing number of features at each round. Then based on the accuracy degradation, a prediction of the same or different author is made. Similarly, Koppel et al. (2011) keep score of how often each author is predicted after random subsets of features are selected, and then make a final prediction based on these scores, dubbed the imposter's method, and Bevendorff et al. (2019) use oversampling with this method to deal with short texts.

Seroussi et al. (2011) use Latent Dirichlet Allocation (LDA), comparing the distance between text representations to determine authorship. They find that this topic-modeling approach can be competitive with the imposter's method while requiring less computation. Seroussi et al. (2014) expand on this topic model approach, and while they present good results on the PAN'11 dataset, the performance of the topic modeling approaches lags behind the best methods. Zhang et al. (2018) introduce a high-performing method that leverages sentence syntax trees and character n-grams as input to a CNN. Saedi and Dras (2021) also presents good results with CNN models, but Ordoñez et al. (2020) indicate that these CNN methods are no longer competitive.

Summary Statistics While older methods focused on small sets of summary statistics, more modern methods are able to combine all of these into a single model. Weerasinghe et al. (2021) provide the best example of this, calculating a plethora of hand-crafted features and Ngrams for each document (distribution of word lengths, hapax-legomena, Maas' a^2, Herdan's V_m, and more). The authors take the difference between these large feature vectors for two texts and then train a logistic regression classifier to predict if the texts were written by the same author or not. Despite its simplicity, this method performs well.

Co-occurance Graphs Arun et al. (2009) construct a graph that represents a text based on the stopwords (nodes) and the distance between them (edge weights). Then to compare the two texts, their graphs are compared using the Kullback-Leibler (KL) divergence. Embarcadero-Ruiz et al. (2022) also construct a graph for each text but instead represent each node as a [word, POS_tag] tuple, and each vertex indicates adjacency frequency. After the graph is created for each text, it is encoded into a one-hot representation and used as input to a LEConv layer. After pooling, the absolute difference between the two document representations is passed through a fully connected network for final scoring.

Embedding Based

Char Embedding Bagnall (2015) use a character-level recurrent neural network (RNN) for authorship verification by sharing the RNN model across all authors but training a different head for each author in the dataset. To classify authors, they calculate the probability that each text was written by each author, predicting the author with the highest probability. Ruder et al. (2016) use both CNNs to embed characters and words for AA. Their results show that the character-based method outperforms the word-based approach across several datasets. Compression-based methods, which leverage a compression algorithm (such as ZIP, RAR, etc.) to build text representations which are then compared with a distance metric, fall into this category as well (Halvani et al., 2017).

Word Embedding Bönninghoff et al. (2019) leverage the Fasttext pre-trained word embeddings, concatenated with a learned CNN character embedding, as part of the input to a bi-directional Long Short Term Memory (BiLSTM) network. That output is then used as input to another network to produce a final document embedding. This neural network structure runs in parallel for two documents (i.e. as a Siamese network Koch et al., 2015), and then optimized according to the contrastive loss function. This method was introduced by Bönninghoff et al. (2019), and then later modified to include Bayes factor scoring on the output by Boenninghoff et al. (2020), and by Bönninghoff et al. (2021) to include an uncertainty adaptation layer for defining non-responses. This was the highest performing method at the PAN20 and PAN21 competitions (Kestemont et al., 2021a).

Jafariakinabad and Hua (2022) build the equivalent of pre-trained word embeddings but for sentence structure (i.e. GloVe-like embeddings that map sentences with a similar structure close together but are agnostic of their meaning), by using the CoreNLP parse-tree and a traditional word-embedded sentence as input to two identical but separate BiLSTMs, and optimize via contrastive loss. The authors also compare against prior work (Jafariakinabad and Hua, 2019) which embeds the POS-tags along with the word embeddings instead of using their custom structural embedding network, showing slight improvement and improved efficiency. CNN's have also been explored given word embeddings as input (Hitschler et al., 2018; Shrestha et al., 2017; Ruder et al., 2016), yet their results are not among the highest.

Transformers Rivera-Soto et al. (2021) build universal representations for AA and AV by exploring the zero-shot transferability of different methods between three different datasets. The authors train a Siamese BERT model (Reimers and Gurevych, 2019b) on one dataset and then test the performance on another without updating. Unfortunately, the results seem to indicate more about the underlying datasets then the ability of these models to uncover a universal authorship representation. Manolache et al. (2021) also explore the applicability of BERT to AA by using BERT embeddings as the feature set for the unmasking method. Comparing this to Siamese BERT, Character BERT (El Boukkouri et al., 2020), and BERT for classification, they find that simple fine-tuning outperforms the more complicated unmasking setup.

Following Bagnall (2015), Barlas and Stamatatos (2020) approach the AA problem by using

	CCAT50	CMCC	Guardian	IMDb62	Blogs50	PAN20	GutenburgAA	Average
Ngram_A	**76.68**	**86.51**	**100**	**98.81**	72.28	**43.52**	57.69	**76.50**
PPM_A	69.36	62.30	86.28	95.90	72.16	—	—	55.14
BERT_A	65.72	60.32	84.23	**98.80**	**74.95**	23.83	**59.11**	66.71
pALM_A	63.36	54.76	66.67	—	—	—	—	26.40

Table 3.2: Macro-accuracy (%) of the authorship attribution models. The "Average"column represents the average macro-accuracy of each model across all datasets in this table, where — entries are counted as 0%.

a shared language model with a different network head for each author. They then compare different shared language model architectures (RNN, BERT, GPT2, ULMFiT, and ELMo), finding that pretrained language models improve the performance of the original RNN architecture. However, the results are all from the small CMCC corpus. Tyo et al. (2021) use a Siamese BERT setup with triplet loss and hard-negative mining for training. Futrzynski (2021) concatenate 28 tokens from each text and then use BERT's `[CLS]` output token for author classification. Peng et al. (2021) concatenate 256 tokens from each text to produce a 512 token input for BERT, and then after pooling use linear layers for same/different author prediction. They repeat this 30 times, sampling different sections of the input texts, and then average over the 30 predictions for final classification.

Feature and Embedding Based

Fabien et al. (2020) explore the applicability of BERT to authorship attribution. They combine the output of BERT with summary statistics via a logistic regression classifier, but find that the summary statistics did not boost performance.

3.3 The VALLA Benchmark

In 1440, Lorenzo Valla proved that the *Donation of Constantine* (where Constantine I gave the whole of the Western Roman Empire to the Roman Catholic Church) was a forgery, using word choice and other vernacular stylistic choices as evidence (Valla, 1922). Inspired by this influential use of AA, we introduce VALLA: A standardized benchmark for authorship attribution and verification.[2] VALLA includes all datasets in Table 3.1, along with others from prior literature (Klimt and Yang, 2004; Manolache et al., 2022; Overdorf and Greenstadt, 2016; Altakrori et al., 2021), with standardized splits, cross-topic/cross-genre/unique author test sets, and usable in either AA or AV formulation. VALLA also includes five method implementations, and we use the

[2]Valla can be found here: https://github.com/JacobTyo/Valla

subscript "A" or "V" to distinguish between the attribution and verification model formulations respectively.

Ngram Being the best performing method in Altakrori et al. (2021), Murauer and Specht (2021), Bischoff et al. (2020), and Stamatatos (2018), this method creates character Ngram, part-of-speech Ngram, and summary statistics for use as input to an ensemble of logistic regression classifiers. For use in the AV setting, we follow Weerasinghe et al. (2021) by using the difference between the Ngram feature vectors of two texts as input to the logistic regression classifier.

PPM Originally developed in Teahan and Harper (2003) and best performing in Neal et al. (2017), this method uses the prediction by partial matching (PPM) compression model (a variant of PPM is used in the RAR compression software) to compute a character-based language model for each author (Halvani and Graner, 2018), and then the cross-entropy between a test text and each author model is calculated. For use in an AV setup, one text is used to create a model and then the cross-entropy is calculated on the second text.

BERT With the highest reported performance on the AA dataset Blogs50 (Fabien et al., 2020) and the most parameters (over 110 million), this method combines a BERT pre-trained language model with a dense layer for classification. For evaluation, we chunk the evaluation text into non-overlapping sets of 512 tokens and take the majority vote of the predictions. For use in the AV setup, the BERT model is used as the base for a Siamese network and trained with contrastive loss (Tyo et al., 2021). For evaluation in the AV setup, we chunk two texts into K sets of 512 stratified tokens (such that the first 512 tokens of each text are compared, the second grouping is compared, etc.), and then take the majority vote of the K predictions.

pALM The best-performing model in Barlas and Stamatatos (2020) was another variation on BERT where a different head was learned on top of the BERT language model for each known author. We refer to this method as the per-Author Language Model (pALM). To classify a text, it is passed through the model for each author, and then the author model with the lowest perplexity on the text is predicted. This is only used in AA formulations as in AV we would have only a single text to train a network head with.

HLSTM Originally introduced by Bönninghoff et al. (2019), this method leverages a hierarchical BiLSTM setup with Fasttext word embeddings and a custom word embedding learned using a character level CNN, as input to a Siamese network. This was the winning method at PAN20 and PAN21 Kestemont et al., 2021a and is only used in AV formulations. While this can be modified to work in AA, we follow prior work and use it only for AV.

All of these methods fall into two categories: those that predict an author class, and those that predict text similarity. The methods that predict an author class (whether via logistic regression, dense layer, etc.) need no post-processing. However, methods that predict similarity need post-processing both for AA and AV problems. For AA, we build an *author profile* by randomly selecting 10 texts from each author and averaging their embeddings together. Then we can compare the unknown texts to each author profile and predict the author that is most similar (in euclidean space). For AV, we directly compare the text representations (again using

euclidean distance) and then define a hard threshold based on a grid search on the evaluation set (although for computing AUC this threshold is irrelevant).

3.4 Experiments and Discussion

All experiments were carried out on 8 V100 GPUs and consumed over 5,000 GPU hours. We optimized for hyperparameters on the validation set via random search, and report all values in the VALLA codebase. All results reported are from a single run that uses the best hyperparameters and is trained until there was no improvement for 2 epochs.

3.4.1 The State-of-the-Art in Authorship Attribution

We start by determining model performance on authorship attribution: given data that is directly attributable to a specific author, learn to classify the work of each author well (macro-accuracy). After evaluating all methods in VALLA on the AA datasets listed in Table 3.1, we find that the traditional Ngram method is the highest performing on average as detailed in Table 3.2. However, we do see that the BERT_{A} model closes the gap on (and can even exceed) the performance of the Ngram_{A} method as the size of the training set increases. This correlation does not hold on the PAN20 dataset, where the best performing model is still Ngram_{A}. This indicates that the state-of-the-art AA method is dependent upon the number of words per author available. While we do not provide a detailed analysis of the data requirements of each method, our results roughly indicate that Ngram_{A} is the method of choice for datasets with less than $50,000$ words per author, while BERT_{A} is the state-of-the-art method for datasets with over $100,000$ words per author. PPM_{A} is simple to tune due to few hyperparameters, but it is both a low performer and it scales poorly to large datasets (rendering it unusable on the PAN20 and Gutenberg datasets). pALM_{A} is the lowest performing method tested, is expensive to train, and scales poorly, so we did not get results on the larger datasets.

The macro-accuracy of BERT_{A} on the IMDb62 and Blogs50 datasets presents a new state of the art, while defining the initial performance marks on the GutenbergAA and PAN20 datasets.[3] The performance on the Blogs50 dataset requires a bit more analysis due to our filtering of duplicates in the dataset. As a better comparison to prior reported performance, we first explore the performance of BERT_{A} on the Blogs50 dataset *without* the filtering, and achieve a macro-accuracy of 64.3%. This represents the state-of-the-art accuracy on a version of the dataset more comparable with prior work (despite its issues) but indicates the strength of the result reported in Table 3.2.

Our results on the Guardian and CMCC datasets are hard to compare to prior work due to

[3]These are initial results because the PAN20 competition was formulated as an AV problem, whereas here we use the AA formulation

	CMCC $\times_t$	CMCC $\times_g$	Guard $\times_t$	Guard $\times_g$
Ngram_A	**82.54**	**84.13**	**86.92**	**87.22**
PPM_A	52.38	57.14	69.23	72.08
BERT_A	49.21	45.24	75.64	75.56
pALM_A	57.14	46.03	61.79	47.22

Table 3.3: Macro-accuracy (%) of the authorship attribution models on domain shifted AA tests sets. $\times_t$ represents cross-topic and $\times_g$ represents cross-genre.

	PAN21	AmaAV	BlogAV	GutAV
Ngram_V	**0.9719**	0.7742	0.5410	0.8741
PPM_V	0.7917	0.6492	0.6230	0.8508
BERT_V	0.9709	**0.8943**	**0.9201**	**0.9624**
HLSTM_V	0.9693	0.8734	0.8580	0.9147

Table 3.4: AUC of the AV models on the selected AV datasets.

the previously mentioned standardization issues, most notably a i.i.d. split has not been used in prior work. The CCAT50 dataset, on the other hand, is directly comparable to prior work. Currently, we show best performing model as the Ngram. However, Jafariakinabad and Hua (2022) report the accuracy of a CNN that takes the syntactic tree of a sentence as input as 83.2% which is better than what we were able to achieve.[4]

3.4.2 The State-of-the-Art in Authorship Attribution under Domain Shift

While dealing with domain shift is an open problem, exploration of domain shift in AA and AV settings is common, even if not explicitly recognized. Table 3.3 examines the performance of the same AA models but focuses on the cross-topic and cross-genre test sets of the CMCC and Guardian datasets. In other terms, the topic (cross-topic) or genre (cross-genre) of the training and test sets are different, therefore giving a lens into how general the models can under such iid violations. Just as in the i.i.d. setting, the Ngram_A method dominates in all scenarios. It should be noted that all datasets used in this domain shift scenario are small, so we cannot verify that the BERT_A method would begin to dominate as the number of words per author increases. We leave the exploration of domain shift performance on larger datasets to future work, although we expect that the BERT_A model would begin to outperform Ngram_A.

	C50	CM	Guard	I62	B50
HLSTM_V	4.56	8.33	27.59	37.82	57.49
(P)HLSTM_V	13.36	16.27	38.97	59.47	11.34
BERT_V	48.64	35.75	27.82	76.62	60.72
(P)BERT_V	56.80	40.87	61.41	73.17	67.21
BERT_A	**65.72**	**60.32**	**84.23**	**98.80**	**74.95**

Table 3.5: Macro-accuracy (%) of the AV models on AA datasets. The (P) indicates that the model was pretrained on the PAN20 training set before fine-tuned on the corresponding dataset. Here we use the following abbreviations: C50 (CCAT50), CM (CMCC), Guard (Guardian), I62 (IMDb62), B50 (Blogs50).

3.4.3 The State-of-the-Art in Authorship Verification

Now we determine model performance on authorship verification: given two texts, determine if they were written by the same author or not. Keeping in line with prior work, the distinction between domain shifted datasets is less clear when formulated as an AV problem. The PAN21 test set is comprised of authors that do not appear in the training set. However the remainder of the datasets (AmazonAV, BlogsAV, and GutenbergAV) are all iid dataset splits. Table 3.4 details the performance of the AV methods on selected AV datasets.

While we saw the Ngram_A method dominating on most AA datasets, here we see that the deep learning-based HLSTM_V and BERT_V methods attain the highest AUC across the board. However, in AV there are only two classes (same and different author), and therefore all of the datasets have a very large number of words per class (vs classes with limited data in AA). Seemingly because of this key difference, AV formulations tend to be more effective for training deep learning methods.

3.4.4 Comparing AA and AV methods

Despite the prominence of comments indicating how AV is the fundamental problem of AA, there is no evidence of how well their performance actually transfers. Table 3.5 shows the performance of LSTM_V and BERT_V on the i.i.d. AA datasets, both when trained only on the dataset as well as starting from a pretrained version of the models (the PAN20 training set was used for pretraining). Here, and in Table 3.6 for the $\times_t$ and $\times_g$ settings, we see notably lower performance than what was obtained by the AA methods. [5]

[4]CCAT50 is a balanced dataset, so the macro-accuracy and accuracy are equal.

[5]We note that the lower performance of the pretrained H-LSTM on Blogs50 than its non-pretrained version is due to the vocabulary selection. This method chooses its vocabulary based on the pretraining corpus, causing transfer issues.

	CMCC $\times_t$	CMCC $\times_g$	Guard $\times_t$	Guard $\times_g$
HLSTM_V	7.94	3.18	19.23	23.33
(P)HLSTM_V	9.52	5.56	40.00	31.53
BERT_V	28.85	13.49	42.31	46.53
(P)BERT_V	33.33	19.05	43.33	54.72
BERT_A	**49.21**	**45.24**	**75.64**	**75.56**

Table 3.6: Macro-accuracy (%) of the authorship verification models on the domain shift AA datasets, where $\times_t$ represents cross-topic and $\times_g$ represents cross-genre. The (P) indicates that the model was pretrained on the PAN20 training set before fine-tuned on the corresponding dataset.

Metric (Formulation)	AUC (AV)	Acc (AV)	Mac-Acc (AA)
BERT_V	0.9229	82.33	67.21
BERT_V w/HNM	**0.9276**	**82.72**	**72.42**

Table 3.7: This table compares the performance of the same model (BERT_V), on the same data (Blogs50), just formulated in different ways, using different performance metrics (column header). w/HNM represents training with hard negative mining.

Hard-Negative Mining We find that AV methods do not necessarily perform well under an AA formulation. To correctly classify a text in the AA setting, a model must make harder comparisons (i.e., compare one text to all others, therefore it will encounter the hardest comparison), whereas an AV setting is strictly easier as it must compare to only a single text. This interpretation motivates the exploration of using hard-negative mining (updating a model during training only on the hardest examples in each batch) for improving the transferability of AV methods to AA problems.

In this section we take a single model (BERT_V) and train two versions of it: one with the contrastive loss and one using triplet loss with batch hard negative mining (specifically the per-batch hard negative mining methodology used in Hermans et al. (2017)). Table 3.7 details these results, showing two key findings. The first is that high AV AUC does not indicate high AA macro-accuracy, and the second is that training an AV method with hard negative mining has little effect on its AV AUC but drastically improves its AA macro-accuracy.

3.5 Conclusion

After a survey of the AA and AV landscapes, we present Valla: an open-source dataset and metric standardization benchmark, complete with implementations of all methods used herein. Using Valla, we present an extensive evaluation of AA and AV methods in a wide variety of common formulations. We achieve a new state-of-the-art macro-accuracy on the IMDb62 (98.81%) and Blogs50 (74.95%) datasets and provide benchmark results on the other datasets.

Our results show that the AV problem formulation is more effective for training deep models. After showing that the high-performing BERT_V does not perform competitively in AA problems, we explore the effect of hard-negative mining on its performance and find that with no degradation in AV performance, it improves the AA macro-accuracy of BERT_V by over 5%, making it a competitive method in the AA formulation. We hope that Valla makes future work in AA and AV more easily approached, and more easily comparable.

3.6 Risks and Limitations

The main risks associated with the development and refinement of AA and AV methods is their misuse. The power to accurately attribute a piece of text to its author holds profound implications, both positive and negative, that warrant careful consideration.

From a privacy perspective, an individual's right to anonymity could be compromised by the misuse of AA and AV methods. While in some circumstances the uncovering of an author's identity is beneficial, such as in forensics or in verifying the authenticity of historical documents, the same technology could also be exploited to unmask authors who wish to remain anonymous for personal, political, or safety reasons.

In this work, we evaluate only on the English language. Furthermore, substantial computational resources were used (over 5,000 GPU hours on V100s). Despite this large amount of compute, after extensive hyperparameter searching, we were only able to get a single run to report metrics on and leave understanding more about the distribution of these results to future work. Both a qualitative analysis, and evaluation of the latest release of large language models are also left to future work.

Chapter 4

Number Detection and Recognition on Off-Road Racers

This paper introduces the Motorcycle Racer Number Dataset (MRND), a new challenging dataset for optical character recognition (OCR) research. MRND contains 2,411 images from professional motorsports photographers' which depict motorcycle racers in off-road competitions. The images exhibit a wide variety of factors that make OCR difficult, including mud occlusions, motion blur, non-standard fonts, glare, complex backgrounds, etc. The dataset has 5,578 manually annotated bounding boxes around visible motorcycle numbers, along with transcribed digits and letters. Extensive experiments benchmark leading OCR algorithms and reveal poor performance on MRND compared to existing datasets. Analysis exposes substantial room for improvement, motivating novel techniques tailored to this domain's unique challenges that are not present in existing OCR datasets. MRND represents an important new benchmark to drive innovation in real-world OCR capabilities. The authors hope the community will build upon this dataset and baseline experiments to make progress on the open problem of robustly recognizing text in unconstrained natural environments.

4.1 Introduction

Optical character recognition (OCR) is a well-studied task in computer vision with immense practical utility. There are many widely deployed systems that require detecting and recognizing textual information from visual data. Thanks to developments in deep learning techniques combined with large annotated datasets, models can now accurately detect and recognize text in images across many languages, contexts and visual domains. Throughout much of its development, research and datasets in OCR have focused on standardized fonts in structured environments, such as typed documents, road signs, and license plates, and OCR systems developed

Figure 4.1: Detecting and recognizing numbers on motorcycles at the start of a race. The top image displays the detected text from a state-of-the-art off-the-shelf OCR model - many of the numbers are not detected or not recognized (bounding boxes with no text prediction). The bottom image displays the detected text from the same model which was further fine-tuned on RnD.

under such controlled conditions are fairly robust and can produce accurate predictions within their corresponding domain Appalaraju et al., 2021; Shashirangana et al., 2020; Netzer et al., 2011.

A much more challenging, but much more versatile setting is recognizing text in unstructured and natural settings. However, recognizing text "in the wild" with unconstrained fonts, orientations, layouts, and contexts remains an open challenge Chen et al., 2021. While it is possible to steer the OCR system to be more robust towards particular settings (e.g. poor lighting) by collecting and annotating data exposed to such conditions, in reality, a natural scene could present a myriad of diverse conditions which can undermine the system's ability to produce accurate text predictions. Furthermore, new domains emerge where current OCR methods struggle due to unique factors previously unseen in existing datasets.

One domain that presents a wide variety of challenging conditions for OCR is recognizing the racer numbers on motorcycles and all-terrain vehicles (ATVs) during off-road racing events (collectively referred to as *motorcycles* in this paper). Racer numbers, which can be used to

identify the racer, are affixed on various locations of each racer and their vehicle. Accurate OCR for racer numbers can enable various useful applications, such as tracking race standings and automated analytics. However, due to the off-road nature of these events, the numbers inevitably exhibit a combination of mud occlusions, non-standard layouts, complex backgrounds, glare, and heavy motion blur. Each of these conditions in isolation presents a major challenge for OCR, and their combination makes this an even more difficult task. Further, to the authors' best knowledge, there exists no public dataset which can support research to tackle these challenges.

To address this gap, we introduce the off-road motorcycle Racer number Dataset (RnD). RnD contains 2,411 images sampled from 16 professional motorsports photographers across 50 different off-road events. The images exhibit the unique challenges of this domain: mud covering numbers, scratches and dirt obfuscating digits, heavy shadows and glare from uncontrolled outdoor lighting, complex backgrounds of other vehicles, bystanders, trees, and terrain, motion blur from rapid maneuvers, large variations in racer number size and location on motorcycles, and various fonts and colors chosen by each racer.

The images are annotated with polygons around every visible motorcycle number along with the transcribed sequence of digits and letters. Only racer identifying texts were annotated. The images were sourced from real racing competitions which span diverse track conditions, weather, lighting, bike types, and racer gear.

The rest of this paper is structured as follows. We first discuss the dataset contents and highlight the domain gaps from existing OCR datasets. We detail the annotation protocol tailored to this domain. We then benchmark leading OCR algorithms to establish baseline accuracy on RnD. The experiments reveal substantial room for improvement, which motivates further research into techniques that can robustly handle mud occlusion, and rapidly evolving perspectives. Our dataset provides the imagery to support developing and evaluating such advances in OCR.

The main contributions are:

- RnD: a off-road motorcycle Racer number Dataset containing 2,411 images with 5,578 labeled numbers sampled from professional photographers at 50 distinct off-road races. To our knowledge, this is the first large-scale dataset focused on recognizing racer numbers in off-road motorsports imagery.
- A rigorous benchmark of generic state-of-the-art OCR models, revealing poor accuracy on RnD and substantial room for innovation.
- Experiments comparing off-the-shelf and fine-tuning strategies. Even the best fine-tuned models fall short.
- Qualitative analysis of prediction errors which provides insights into failure modes to guide future research directions.

We hope RnD and our initial experiments will catalyze innovation in real-world text recognition capabilities. Robust reading of racer numbers has potential applications in race analytics,

timing systems, media broadcasts, and more. Our work reveals this as an open research problem necessitating domain-targeted techniques.

4.2 Related Work

Text detection and recognition in images is a classic computer vision task. Early traditional methods relied on sliding windows, connected components, and handcrafted features like HOG Wang and Belongie, 2010. With the advent of deep learning, convolutional and recurrent neural networks now dominate scene text recognition pipelines Chen et al., 2021. Models leverage large annotated datasets to learn powerful representations tuned for text detection and recognition in a specific domain.

Many datasets and competitions have driven progress in general OCR. These include ICDAR Karatzas et al., 2013, COCO-Text Lin et al., 2014, and Street View Text Wang et al., 2011a. Popular detection models build on Region Proposal Networks and include CTPN Tian et al., 2016, EAST Zhou et al., 2017, and Craft Baek et al., 2019. Recognition is often achieved via CNN + RNN architectures like CRNN Wang et al., 2019 or transformer networks like ASTER Shi et al., 2019. More recent state-of-the-art methods utilize pre-trained vision models like ViTSTR Atienza, 2021, PARSeq Bautista and Atienza, 2022, CLIP4STR Zhao et al., 2023, and DeepSolo Ye et al., 2023. However, most OCR research targets images of documents, signs, or web images. While many of these works aim to go beyond structured settings (e.g.images of documents, signs, or web images) and address the task of "robust reading", i.e. OCR in incidental or real scenes, recognizing text "in the wild" with few assumptions remains an open challenge Chen et al., 2021. Furthermore, domain gaps exist where current methods fail on specialized applications. Our work focuses on one such gap - recognizing racer numbers in motorsports.

A few prior works address detecting and recognizing the license plates on vehicles Ap et al., 2020; Laroca et al., 2018; Chen et al., 2019b; Lee et al., 2019b; Silva and Jung, 2018; Quang et al., 2022; Laroca et al., 2021. Some have focused specifically on street motorcycle number plates Kulkarni et al., 2018; Sathe et al., 2022; Sanjana et al., 2021; Lee et al., 2004. All of these efforts use data gathered from some form of street camera, which are placed in strategic locations with recognizing license plates specifically in mind. In contrast, our dataset is gathered from professional motorsport photographers focused on capturing the most aesthetically pleasing photograph of each racer. Furthermore, existing datasets have standardized plates which differ greatly from the diverse layouts and occlusions of off-road motorcycle numbers. Street motorcycle plates exhibit consistency in position and appearance, unlike the numbers encountered during off-road competitions. The conditions during races also introduce and exacerbate factors like motion blur, mud occlusion, glare, and shaky cameras not prevalent in street imagery. RnD provides novel real-world imagery to push OCR capabilities.

The most relevant prior domain is recognizing runner bib numbers in marathon images Shiv-

(a) (b) (c) (d)

Figure 4.2: Common locations and variations of racer numbers. (a) Numbers can be seen on the hand guards, and vegetation close to the photographer makes for a new sort of occlusions. (b) The front number, side number, and helmet number are all different. (c) Numbers can be on the back of racer's jerseys. (d) Different front and side numbers.

akumara et al., 2017; Ben-Ami et al., 2012; Boonsim, 2018; Kamlesh et al., 2017. This shares similarities, but runner bibs provide more spatial and appearance consistency than motorcycle racing numbers. Datasets like TGCRBNW Hernández-Carrascosa et al., 2021 exhibit some motion blur and night racing, but do not contain the mud, vehicle occlusion and diversity of layouts seen in motorsports.

Number recognition has also been studied in other sports - football Yamamoto et al., 2013; Bhargavi et al., 2022, soccer Gerke et al., 2015; Gerke et al., 2017; Šaric et al., 2008; Diop et al., 2022; Alhejaily et al., 2023, basketball Ahammed, 2018, track and field Messelodi and Modena, 2013, and more Liu and Bhanu, 2019; Nag et al., 2019; Vats et al., 2021; Wrońska et al., 2017. However, most focus on jersey numbers in commercial broadcast footage rather than track/field-side imagery. Existing sports datasets offer limited diversity and size. To our knowledge, RnD represents the largest, most varied collection of motorsports numbers in natural contexts.

In summary, prior work has made great progress in OCR for documents, signs, and other domains, but real-world applications like recognizing racers in off-road competitions remain extremely challenging due to domain gaps in current data. RnD provides novel imagery to spur advances in OCR for motorsports. Our benchmark experiments expose substantial room for improvement using this data.

(a) (b) (c) (d)

Figure 4.3: Examples of some difficult, but not muddy, images. (a) Two separate numbers are on the front of the motorcycle, a smaller number overlapping a bigger number. Furthermore, half of the number plate is not legible due to glare. (b) The front-brake cable overlaps the number. (c) A racer is crashing, resulting in contrived number orientations. (d) Shadows cast from trees cause difficult lighting conditions.

4.3 Dataset

The off-road motorcycle Racer number Dataset (RnD)[1] is comprised of 2,411 images gathered from the off-road photography platform PerformancePhoto.co. Each image depicts motorcycle racers engaged in competitive events, with visible racer numbers on themselves and their motorcycles. The dataset includes bounding box annotations and transcriptions from over 50 different off-road motorcycle and ATV races. The races cover various track conditions, weather, and lighting. The images were captured by 16 different professional photographers using a wide range of high-end cameras.

Racers can have anywhere from one to as many as 20 numbers located on their body and motorcycle. The common locations for a number include the front and sides of the motorcycle, on the cheeks of the racer's helmet, and on the back of the racers jersey. However, in rare cases, numbers can also be seen on the wheels and handguards. The numbers on a single racer and vehicle do not need to all be the same number. Commonly, the numbers on the helmet do not match the numbers on the motorcycle, and the number on the front of the motorcycle does not need to match the number on the side. It is also common for numbers to only be present on the racer, but not on the motorcycle. Figure 4.2 highlights some of these examples.

In RnD, there is a total of 5,578 racer number annotations. The numbers can span from 1 to 5 characters in length, optionally including alphabetical characters (e.g., adding a letter to the end of a number is a common modifier - for convenience, we still refer to all of these as *numbers*). 6% of the dataset includes numbers that have alphabetical characters in addition to the numerics. The dataset is split randomly into a training and a testing set, with 80% of the

[1]The dataset is available at https://github.com/JacobTyo/SwinTextSpotter.

(a) (b) (c) (d) (e)

Figure 4.4: Mud poses the most significant challenge to effective OCR in this domain. (a) Not only is the racer in an odd pose, but the number is also occluded in sticky mud. (b) The racer is covered in wet mud, posing a different, although more managable, type of mud occlusion. (c) Mud occlusions in sandy environments again poses new types of occlusions. (d) An extreme example of sticky mud completely obscuring all details about a racers number. (e) Generic example of the most commonly seen type of mud occlusion.

images in the training set.

4.3.1 Annotation Process

Only the racer numbers were annotated instead of all visible text by one of the authors. All visible racer numbers were tightly bounded by a polygon (i.e. the bounding box), and each polygon is tagged with the characters contained within (i.e. the number). If a character was ambiguous or unclear, it was labeled with a '#' symbol. Only the humanly identifiable text was transcribed. Any racer numbers that were fully occluded or too blurry to discern were not annotated.

The transcription task was restricted to only use the context of each individual bounded region. The full image context could not be used to infer ambiguous numbers based on other instances of that racer's number elsewhere on the motorcycle. This simulates the local context available to optical character recognition models.

4.3.2 Dataset Analysis

Figure 4.3 highlights some of the challenging factors present in this dataset. Lighting conditions vary from extremely bright to extremely dark (including night races). Figure 4.3a gives an example of glare that is common in a field with exposure to sunlight (8% of images), and Figure 4.3d shows the complications that the forest can cause on lighting conditions (7% of images). Not only are there occlusions typical of other datasets such as trees or other racers blocking

Table 4.1: Comparison of the text detection and recognition performance on the RnD test set using off-the-shelf versus fine-tuned state-of-the-art OCR models. Precision, recall, and F1 score are reported for both detection (Det-P, Det-R, Det-F1) and end-to-end recognition (E2E-P, E2E-R, E2E-F1). The off-the-shelf versions achieve very low scores, while fine-tuning improves results substantially. However, even fine-tuned models fall short of real-world viability, with the best YAMTS model obtaining only 0.527 end-to-end F1 score. This highlights significant room for improvement using domain-targeted techniques and data such as RnD.

Model		Det-P	Det-R	Det-F1	E2E-P	E2E-R	E2E-F1
OTS	SwinTS	0.195	0.287	0.232	0.101	0.148	0.120
	YAMTS	0.192	0.491	0.276	0.106	0.244	0.148
FT	SwinTS	0.810	0.673	0.734	0.513	**0.415**	0.459
	YAMTS	**0.847**	**0.715**	**0.775**	**0.758**	0.404	**0.527**

the view, but we are also presented with extremely challenging cases where a smaller number is placed over top of a bigger number (See Figure 4.3a). In such cases, we label every number we can properly identify. Furthermore, as shown by the front brake cable in Figure 4.3b, some motorcycles have components that pass in front of the number plate. Finally, orientation of the numbers vary greatly, not only due to the nature of motorcycles (i.e. they must be leaned over to turn corners), but also in cases such as crashes, as shown in Figure 4.3c.

The most unique aspect of this dataset is a new type of occlusion: mud. Mud is frequently encountered in off-road racing, and Figure 4.4 gives examples ranging from light to extreme (44% of images). In the worst of cases, it is impossible to detect any racer numbers (Figure 4.4d). However, in many cases, humans are still able to accurately complete this task.

4.4 Experiments

We conducted experiments to benchmark the performance of modern OCR methods on the RnD. Our goals here are twofold: 1) establish baseline results on this new domain, and 2) analyze where current algorithms fail. Four NVIDIA Tesla V100 GPUs were used for these experiments. Hyperparmeter searching was performed

4.4.1 Models

Our experiments leverage two state-of-the-art scene text spotting models:

- **YAMTS**: Yet Another Mask Text Spotter Krylov et al., 2021

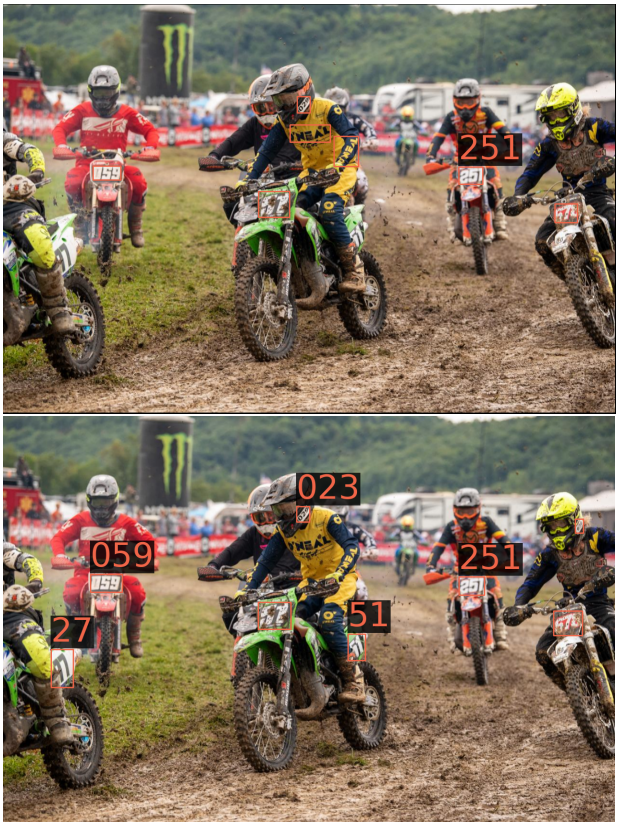


Figure 4.5: Example showcasing model successes and failures on a complex muddy image. The top image shows detected text from the off-the-shelf YAMTS model before fine-tuning, which recognizes only 1 number correctly ("251"). The bottom image displays results from the fine-tuned YAMTS model, which detects all 8 visible numbers but only correctly recognizes 3 of them. This highlights benefits of domain-specific fine-tuning, as the pre-trained model struggles. However, even the fine-tuned model has difficulty accurately recognizing highly degraded text, exposing substantial room for improvement.

YAMTS is a Mask R-CNN-based model with an additional recognition head for end-to-end scene text spotting. A ResNet-50 He et al., 2016 is used for text detection, with a convolutional text encoder and a GRU decoder.

- **SwinTS**: Swin Text Spotter Huang et al., 2022
  The Swin Text Spotter is an end-to-end Transformer-based model that improves detection and recognition synergy through a recognition conversion module. A feature pyramid network is used to decrease the sensitivity to text size, and the recognition conversion model enables joint optimization of the detection and recognition losses.

For both models, we first benchmark their performance on the RnD test set using their published pre-trained weights, which are from training on a large corpus of training data. YAMTS was pretrained on Open Images V5 Kuznetsova et al., 2020; Krylov et al., 2021, ICDAR 2013 Karatzas et al., 2013, ICDAR 2015 Karatzas et al., 2015, ICDAR 2017 Gomez et al.,

Figure 4.6: Example showcasing the fine-tuned model learning to see through mud. The left image depicts the predictions from the off-the-shelf YAMTS model before fine-tuning, which does not recognize any text. The right image displays results from the fine-tuned YAMTS model, which is able to see through the heavy mud occlusion and properly detect and recognize the racer number. This demonstrates improved robustness to real-world mud occlusion after domain-specific fine-tuning.

2017, ICDAR 2019 Zhang et al., 2019, COCO-text Veit et al., 2016, and MSRA-TD500 Yao et al., 2012. SwinTS was pretrained on Curved SynthText Liu et al., 2020b, TotalText Ch'ng and Chan, 2017, ICDAR 2013 Karatzas et al., 2013, and ICDAR-MLT Gomez et al., 2017; Zhang et al., 2019.

Afterwards, we fine-tune these models further on the RnD training set and evaluate their performance again. We first performed a grid search over the learning rate, learning rate schedule, warm-up period, and batch size using the validation set. We found the best setup to be a cosine annealing learning rate schedule with a warm up, using a batch size of 8 images across 4 GPUs, with the random scaling and rotation data augmentations. The learning rate starts at 1e-6 and is then raised to 1e-3 after 1,000 iterations, and then annealed back down to 1e-6 over the remainder of training. These hyperparameters were used to fine-tune the models over 150 epochs. The fine-tuned models are evaluated on the RnD test set.

4.4.2 Evaluation Metrics

Following the standard evaluation protocol Huang et al., 2022; Ye et al., 2023, we report results for both the text detection and end-to-end recognition tasks. For detection, we compute precision, recall, and F1-score, which we denote *Det-P*, *Det-R*, and *Det-F1* respectively. A predicted box was considered a true positive if it overlapped with a ground truth box by at least 50% intersection over union. For end-to-end recognition, we report precision, recall, and F1-score at the sequence level, and we likewise denote these metrics as *E2E-P*, *E2E-R*, and *E2E-F1*. A predicted

(a) (b) (c) (d) (e)

Figure 4.7: Analysis of model performance on mud occluded numbers. (a) Model correctly recognizes front number by ignoring mud. (b) Quad number is recognized but muddy helmet number is missed. (c) Front number is read but very muddy helmet number is missed. (d) Number is detected but misrecognized due to odd position. (e) Two numbers are correctly read but muddy side number is missed.

(a) (b) (c) (d) (e)

Figure 4.8: Analysis of common non-mud failures: (a) Incorrect side number recognition. (b) Overlapping "stacked" numbers confuse the model. (c) A letter is mis-recognized as a number. (d) The letter portion of the racer number is missed. (e) Complex graphics on quad confuse model.

text sequence was considered correct only if it exactly matched the ground truth transcription for the corresponding ground truth box.

4.5 Results and Discussion

Table 4.1 summarizes the quantitative results on the RnD test set. The off-the-shelf SwinTS and YAMTS models, which were pretrained on large generic OCR datasets, achieve poor accuracy. This highlights the substantial domain gap between existing datasets and this new motorsports application. Even state-of-the-art models fail without adaptation to racer numbers.

Fine-tuning the pretrained models on RnD led to major improvements. SwinTS achieved

Figure 4.9: Example showcasing model improvement in rainy conditions. The top image shows detections from the off-the-shelf YAMTS model before fine-tuning, which recognizes only 1 number correctly (“35”). The bottom image displays results from the fine-tuned YAMTS model, which detects all 6 visible numbers and correctly recognizes 5 of them.

0.734 detection F1 and 0.459 end-to-end recognition F1 after fine-tuning. For YAMTS, fine-tuning improved to 0.775 detection and 0.527 recognition F1 scores. However, these fine-tuned results still fall short of requirements for robust real-world deployment.

The experiments reveal substantial room for improvement over state-of-the-art methods on RnD. Neither off-the-shelf nor fine-tuned models achieve sufficient accuracy for motorcycle racing applications, which we detail further in the next section with qualitative analysis. Overall, our quantitative benchmarks establish baseline results to motivate innovative techniques tailored to OCR on muddy vehicles in dynamic outdoor environments.

4.5.1 Performance Among Occlusion

Table 4.2: Performance broken down by occlusion.

Occlusion (%)		Det-P	Det-R	Det-F1	E2E-P	E2E-R	E2E-F1
None (41%)	OTS	0.196	0.568	0.291	0.124	0.330	0.180
	FT	0.880	0.726	0.795	0.826	0.470	0.599
Blur (3%)	OTS	0.231	0.545	0.324	0.140	0.295	0.190
	FT	0.860	0.841	0.851	0.750	0.409	0.529
Shadow (7%)	OTS	0.144	0.536	0.227	0.033	0.107	0.050
	FT	0.875	0.778	0.824	0.769	0.370	0.500
Mud (44%)	OTS	0.194	0.389	0.259	0.086	0.152	0.110
	FT	0.811	0.718	0.761	0.681	0.359	0.470
Glare (8%)	OTS	0.162	0.547	0.250	0.052	0.156	0.078
	FT	0.787	0.686	0.733	0.519	0.200	0.289
Dust (2%)	OTS	0.173	0.310	0.222	0.113	0.190	0.142
	FT	0.925	0.638	0.755	0.833	0.259	0.395

We further analyzed model performance on the RnD test set when numbers were occluded by different factors. Note that a single image can contain multiple occlusions (i.e. it can be dusty and have glare, or it can be blurry and muddy, etc.). Table 4.2 breaks down the detection and recognition results on images with no occlusion, motion blur, shadows, mud, glare, and dust.

Mud occlusion was the most prevalent, accounting for 44% of the test data. Both off-the-shelf and fine-tuned models struggled with heavy mud. The fine-tuned model improved over the off-the-shelf version, achieving 0.761 detection F1 and 0.470 recognition F1 on muddy images. But this remains far below the 0.795 detection and 0.599 recognition scores attained on non-occluded data. There is substantial room to improve robustness to real-world mud and dirt occlusion.

The fine-tuned model also struggled with glare occlusion, scoring just 0.733 detection F1 and 0.289 recognition F1 on such images. Glare creates low contrast regions that likely hurt feature extraction. Shadows likewise proved challenging, with a 0.824 detection but only 0.500 recognition F1 score after fine-tuning. The changing lighting and hues may degrade recognition.

For motion blur, the fine-tuned model achieved 0.851 detection F1 but 0.529 recognition F1. Blurring degrades the crispness of text features needed for accurate recognition. Surprisingly, the model performed worst on dust occlusion, despite it being visually less severe than mud and glare. This highlights brittleness of vision models to unusual textures.

Overall, the breakdown reveals mud as the primary challenge, but substantial room remains to improve OCR accuracy under real-world conditions like shadows, dust, blur, and glare. Researchers should prioritize occlusions seen in natural operating environments that undermine off-the-shelf models.

4.5.2 Qualitative Analysis

We analyzed model performance on RnD using the fine-tuned YAMTS model, which achieved the highest end-to-end F1 score. The detection confidence threshold was set to 0.65 and the recognition threshold set to 0.45. Figures 4.5-4.9 showcase successes and failures on challenging examples. When side-by-side comparisons are drawn, we compare against the off-the-shelf YAMTS model before fine-tuning.

Figure 4.5 compares the text spotting performance before and after fine-tuning on a photo of the start of a muddy race. The fine-tuned model properly detects all 8 visible numbers, demonstrating capabilities to handle partial mud occlusion. However, it only correctly recognizes 3 of the 8 numbers, highlighting limitations recognizing degraded text. Without fine-tuning, only 1 number is detected, and no numbers are properly recognized, showing benefits of fine-tuning. But substantial challenges remain in muddy conditions.

Figure 4.7 showcase common mud-related successes and failures. In some casese, the fine-tuned models are able to see through mud occlusions to properly recognize the racer number, as shown in Figure 4.7a. However, mud often prevents smaller helmet numbers from being recognized (Fig 4.7b, 4.7c). Odd orientations also confuse models (Fig 4.7d). Overall, heavy mud occlusion remains the biggest challenge. Figure 4.8 reveals other common failures like missing side numbers (Fig 4.8a), overlapping numbers (Fig 4.8b), confusion between letters and numbers (Fig 4.8c), missing letter portions (Fig 4.8d), and distractions from graphics (Fig 4.8e). In summary, the analysis reveals promising capabilities but also exposes key areas for improvement, particularly among extreme mud and small text. Substantial opportunities remain to enhance OCR for this challenging real-world application.

Photos from the beginning of a race are typically the most complex, due to the number of motorcycles in a single image and background clutter. Figure 4.9 again looks at a photo from

the start of a race, but this time in rainy conditions. The top photo highlights the detections of the off-the-shelf model before fine-tuning, where it is able to recognize only a single number properly. However, after fine-tuning, the model is able to properly recognize 5 of the 6 visible numbers.

4.6 Conclusion

In this work, we introduced the off-road motorcycle Racer number Dataset (RnD), a novel challenging real-world dataset to drive advances in optical character recognition. RnD contains 2,411 images exhibiting factors such as mud, motion blur, glare, complex backgrounds, and occlusions that degrade text detection and recognition accuracy. The images were captured by professional motorsports photographers across 50 distinct off-road competitions.

We annotated 5,578 racer numbers with transcriptions and tight bounding boxes. The data exhibits natural diversity in lighting, weather, track conditions, vehicle types, racer gear, and more. To our knowledge, RnD represents the largest, most varied collection of annotated motorsports numbers in unconstrained environments.

We established baseline results on RnD using the state-of-the-art text spotting models, Swin Text Spotter and YAMTS. Off-the-shelf versions pretrained on generic OCR data achieved an end-to-end F1 score around 0.2, highlighting the sizable domain gap. Fine-tuning on RnD improved results but even the best model obtained only 0.527 end-to-end F1, far below practical expectations for real-world use. Through qualitative analysis, we revealed some of the primary factors degrading OCR accuracy on RnD to be heavy mud occlusion, glare, dust, and more. Heavily distorted fonts and unusual orientations also led to several notable mistakes.

Overall, our work exposes motorcycle racer number recognition as an open challenge with unique conditions, and provides a dataset of novel real-world imagery. The experiments establish baseline results using leading methods, quantitatively and quantitatively demonstrating substantial room for improvement on RnD. We hope the community will build upon these initial experiments to make advances on the problem of accurately reading text in unconstrained natural environments.

Chapter 5

Re-Identification of Off-Road Racers

Re-identifying individuals in unconstrained environments remains an open challenge in computer vision. We introduce the Muddy Racer re-IDentification Dataset (MUDD), the first large-scale benchmark for matching identities of motorcycle racers during off-road competitions. MUDD exhibits heavy mud occlusion, motion blurring, complex poses, and extreme lighting conditions previously unseen in existing re-id datasets. We present an annotation methodology incorporating auxiliary information that reduced labeling time by over 65%. We establish benchmark performance using state-of-the-art re-id models including OSNet and ResNet-50. Without fine-tuning, the best models achieve only 33% Rank-1 accuracy. Fine-tuning on MUDD boosts results to 79% Rank-1, but significant room for improvement remains. We analyze the impact of real-world factors including mud, pose, lighting, and more. Our work exposes open problems in re-identifying individuals under extreme conditions. We hope MUDD serves as a diverse and challenging benchmark to spur progress in robust re-id, especially for computer vision applications in emerging sports analytics.

5.1 Introduction

Re-identifying individuals across disjoint camera views is a fundamental task in computer vision. Despite progress, most research assumes controlled capture environments and consistent appearances (Gou et al., 2018). When restricted to such controlled environments, existing solutions do a good job of handling challenges due to occlusion, pose variation, and lighting changes. although further progress is needed Ye et al., 2021b. However, as we show, outside of such controlled environments, current techniques struggle.

In particular, we focus on the task of identifying motorcycle racers during off-road competitions (Figure 5.1) through mud, dirt, trees, and crowds. Here, appearances can change drastically lap-to-lap as mud accumulates or subsequently flies off. Numbered jerseys that could otherwise

Figure 5.1: Motorcycle Racer Re-Identification

be used to easily re-id racers often become obscured by mud, are out of sight of the camera, or get torn. Glare, blurring, and extreme lighting also occur as a single racing event can go from bright fields to deep dark forests. To the best of our knowledge, there exists no public datasets to supports research into robust re-id under such conditions.

To spur progress in addressing these challenges, we introduce the Muddy Racer Re-Identification Dataset (MUDD). MUDD contains $3,906$ images of 150 identities captured over 10 off-road events by 16 professional motorsports photographers. The imagery exhibits heavy mud occlusions, complex poses, distant perspectives, motion blurring, and more. We also present an efficient annotation methodology incorporating detected racer numbers as auxiliary information to generate high-quality identity clusters for manual verification. This improved labeling time by over 65% compared to more simplistic labeling methods.

We establish benchmark performance using state-of-the-art re-id models based on a Omni-Scale CNN Neural Network (Zhou et al., 2019a) and ResNet-50 (He et al., 2016). Without fine-tuning, the best models reach only 22% Rank-1 accuracy. But when fine-tuning is incorporated, the best models reach 79% Rank-1 accuracy. Interestingly, pretraining can be performed with ImageNet data (Deng et al., 2009) to achieve nearly identical performance as pretraining on re-identification (re-id) specific datasets. Despite this increase in performance, a considerable

gap remains between machine and human performance.

Our analysis reveals open problems in handling mud occlusion, appearance changes, poses, resolution, and similar outfits. These factors induce intra-class variation and inter-class similarity that current models fail to robustly distinguish. In summary, we introduce a diverse, challenging dataset exposing the limitations of existing re-id techniques. MUDD provides imagery to drive progress in re-identification amidst uncontrolled real-world conditions.

Our contributions are:

- The MUDD dataset containing diverse imagery to evaluate re-id of off-road racers. To our knowledge, this represents the first large-scale dataset of this emerging application domain.
- A method to improve annotation effectiveness by incorporating auxiliary information during labeling.
- Initial benchmarking of state-of-the-art models, which reveal limitations on this dataset and substantial room for further improvement.
- Analysis of failure cases which provide insights to guide future research on robust re-identification for sports analytics and computer vision broadly.

5.2 Efficient Labeling via Auxiliary Information

One key challenge in constructing re-id datasets is to efficiently group images of the same identity during labeling. Exhaustively labeling identities from scratch can become intractable for a large dataset of images with an unknown number of identities. To assist in this labeling process, images can be clustered into groups using pretrained models, and then manually verified by annotators. However, in constructing MUDD, we found that annotators still spent over 30 minutes on each identity, requiring a more efficient process.

Off-the-shelf re-id models focus on extracting features invariant to nuisance factors like pose, lighting, and blurring while discriminating between identities. However, these features are based on their pretraining dataset and they cannot explicitly leverage domain-specific cues—especially if the domain-specific cues are not available in the pretraining dataset, such as racer numbers. The re-id model treats the image holistically without localizing and recognizing semantic concepts like digits. Therefore, when the models are applied on different image domains, any useful domain-specific cues are not used.

In light of this challenge, we leverage the fact that each identity (i.e. racer) in this dataset is assigned a visible number and we propose directly utilizing this auxiliary information during the clustering and re-id process via a pretrained text detection model (Lyu et al., 2018). This domain knowledge provides strong localization cues to group images with the same numbers. The re-id model alone struggles to consistently spot and match the small digit regions amidst

Table 5.1: MUDD re-id benchmark results comparing off-the-shelf, from scratch, and fine-tuning training strategies. Fine-tuning provides major accuracy gains indicating the importance of transfer learning.

Training	Backbone	R1	R5	R10	mAP
Off-the-shelf	OSNet	0.325	0.522	0.633	0.385
	ResNet-50	0.316	0.510	0.631	0.363
Trained From Scratch	OSNet	0.215 (0.031)	0.485 (0.045)	0.676 (0.036)	0.249 (0.016)
	ResNet-50	0.159 (0.012)	0.416 (0.016)	0.619 (0.283)	0.192 (0.019)
Pretrained on Imagenet	OSNet	0.784 (0.013)	0.942 (0.006)	0.977 (0.005)	0.822 (0.013)
	ResNet-50	0.762 (0.008)	0.944 (0.004)	0.979 (0.003)	0.807 (0.006)
Pretrained on MSMT17	OSNet	0.792 (0.010)	0.945 (0.002)	0.978 (0.002)	**0.829 (0.006)**
	ResNet-50	0.760 (0.023)	0.941 (0.012)	0.977 (0.007)	0.803 (0.024)
Pretrained on DukeMTMC	OSNet	0.789 (0.015)	0.939 (0.003)	97.57 (0.004)	0.826 (0.012)
	ResNet-50	0.786 (0.017)	**0.956 (0.008)**	**0.985 (0.002)**	0.828 (0.011)
Pretrained on Market-1501	OSNet	**0.793 (0.017)**	0.944 (0.006)	0.978 (0.005)	0.827 (0.015)
	ResNet-50	0.781 (0.025)	0.948 (0.014)	0.981 (0.009)	0.823 (0.021)

mud, motion, and variations.

Explicitly guiding search and clustering with the auxiliary numbers, even if noisy, complements the holistic re-id model. Our breadth-first attribute search leverages the domain knowledge to effectively explore the data and retrieve number matches. This creates high-quality initial clusters that seed the depth-first re-id search.

In essence, we get the best of both worlds: domain-driven localization from the auxiliary cues, combined with holistic identity discrimination from the re-id model. The re-id model alone lacks the explicit semantic guidance, resulting in poorer search and clustering. Our hybrid approach better utilizes both domain knowledge and learned representations.

Specifically, to generate ground truth labels for specific racers, we first extract all numbers using a pretrained text detection model (Lyu et al., 2018), and also create a re-id embedding using a pretrained OSNet model. Then we iterate over the following process:

1. Pick a number that was detected more than 10 times and retrieve all images containing it.
2. For each result from Step 1, take the top k nearest neighbors based on the re-id embedding.
3. Combine the results for each search by rank, and present to annotators for manual refinement and verification.

This updated process reduced the average time to verify an identity cluster from over 30 minutes to under 10.

Figure 5.2 shows a proposed cluster from our labeling system. The top section contains all photos where the number 530 was detected. The bottom section shows the most similar images according to the pertained OSNet re-id model. Critically, leveraging the auxiliary number information provides an initial cluster with clean and muddy images of the same racer that can be used as a seed image for a search by the re-id model. Figure 5.3 shows additional results deeper in the ranking.

5.3 The MUDD Dataset

MUDD[1] contains 3906 images capturing 150 identities across 10 different off-road events from the Grand National Cross Country (GNCC) racing series. The events span various track conditions, weather, times of day, and racing formats. Images were captured by 16 professional motorsports photographers using a diverse range of high-end cameras.

We gathered a large library of off-road competition photos from the off-road photography platform PerformancePhoto.co. We used YOLOX (Ge et al., 2021) to detect the bounding boxes for people. An embedding was extracted for each cropped bounding box using the general-use re-id model OmniScaleNet (Zhou et al., 2019a). We leverage a scene text spotter (Lyu et al., 2018) to extract visible racer numbers as auxiliary information to aid our labeling process, as detailed in Section 2. Importantly, the accuracy of the auxiliary models on MUDD data is low. Our scene text spotter has less than 50% end-to-end accuracy. However, as described in Section 2, even low-accuracy auxiliary information can still drastically improve annotation efficiency by enabling effective search and clustering.

We manually labeled all identities, accelerated by our proposed method. Some individuals occur in multiple events, either with very similar outfits or entirely different ones. To simplify training and evaluation, we provide an event ID and treat the same individual across events as different identities.

MUDD contains several major challenges:

- Heavy mud occlusion - Racers accumulate significant mud spatters and caking. This represents a unique occlusion pattern not present in existing re-id datasets.
- Complex poses—Racers exhibit varied poses including leaning, jumping, crashes, and more unseen during regular walking.
- Distance and resolution—Images captured from a distance with small, low-resolution racer crops.
- Dynamic lighting—Outdoor conditions cause glare, shadows, and exposure variations.
- Clothing—Jerseys and numbers that could ease re-id are often obscured by mud, gear, and positioning.

[1]MUDD is available at https://github.com/JacobTyo/MUDD

- Motion blur—Racers maneuver at high speeds causing motion-blurring effects, especially combined with panning cameras.

We divided MUDD into train (80%) and test (20%) sets. There is no identity overlap between the sets. We further divided the train set into a train and validation split also with a 90/10 ratio. The validation set was used for model selection, hyperparameter tuning, and ablation studies. All metrics reported on the held-out test set.

The dataset includes identities under a variety of motorcycle and riding gear. It captures both professional and amateur events across multiple states during the first 7 months of 2023. The diversity of identities, environments, perspectives, and conditions exceeds existing re-id datasets.

5.4 Experiments

We evaluated the performance of models on MUDD in three settings:

- **Off-the-shelf**: Pre-trained state-of-the-art re-id models applied directly to MUDD.
- **Random Initialization**: Models trained from random initialization only on MUDD.
- **Transfer**: Person re-identification pre-trained models fine-tuned on MUDD.

We selected strong open-source implementations of CNN-based architectures, hereafter referred to as OSNet (Zhou et al., 2019a) and ResNet50 (He et al., 2016).

For training, we used triplet loss, and data augmentation of random flips, color jitter, and random crop. Models were optimized using Adam. We tuned hyperparameters like learning rate, batch size, and data augmentation techniques based on the validation set. All models were trained for 100 epochs, using a cosine learning rate schedule with a maximum learning rate of 0.0003. The final performance is reported on the test set at the best checkpoint, and all models were trained on a single NVIDIA 2080Ti GPU. The mean and standard deviation are reported over three random seeds.

5.4.1 Evaluation Metrics

We adopt the standard re-id metrics cumulative matching characteristic (CMC) rank-1, rank-5, rank-10 and mean Average Precision (mAP). CMC measures rank-k accuracy, the probability of the true match appearing in the top k. The mAP metric computes mean average precision across all queries. Both operate directly on the re-id model output.

5.5 Results

Table 5.1 summarizes re-id performance on MUDD.

Off-the-shelf Models Applying pre-trained re-id models directly to MUDD leads to very poor accuracy. The highest Rank-1 is only 32.52% using OSNet pre-trained on Market-1501. This highlights the significant domain gap between existing re-id datasets and MUDD's challenging conditions. Off-the-shelf models fail to generalize.

Training from Scratch Begining with a random initialization, training models directly on MUDD struggles to learn effectively. OSNet is only able to achieve 21.46% Rank-1 accuracy, indicating the difficulty of learning a robust representation from this training data alone.

Fine-tuning Fine-tuning pre-trained models by resuming optimization on MUDD significantly improves accuracy. Fine-tuned OSNet reaches 79.31% Rank-1, over 2.5x higher than off-the-shelf and 3.7x higher than training from scratch. Fine-tuning transfers invariant features and discrimination capabilities from larger source datasets, allowing models to adapt to MUDD despite the limited training data.

Interestingly, models pre-trained on generic ImageNet data perform nearly as well as those pre-trained on re-id specific datasets like Market-1501 after fine-tuning. This indicates MUDD represents a significant domain shift even from existing re-id datasets. The ImageNet features still provide a useful initialization for fine-tuning to this new domain.

Architectures We experimented with two CNN-based architectures: OSNet, specifically designed for re-id tasks Zhou et al., 2019a, and ResNet-50, a general-purpose CNN also commonly used for re-id He et al., 2016. After fine-tuning on MUDD, OSNet achieves slightly higher Rank-1 accuracy (79.31%) than ResNet-50 (78.12%).

This performance gap may stem from OSNet's specialized representations tailored for scale-invariance on people. In contrast, ResNet's more general features still perform competitively, demonstrating the versatility of standard CNNs. Overall, both architectures can adapt to MUDD's domain when fine-tuned, with OSNet's inductive biases providing a small boost. However, substantial room for improvement remains compared to human performance.

Pretraining Datasets We considered models already tailored to the person re-identification task. Starting with models pretrained on one of the re-id datasets of MSMT17 (Wei et al., 2018),

Table 5.2: The performance of the best method when controlling the query and gallery set for muddy images. "No Mud ->Mud" corresponds to the query set containing only clean images, and the gallery set containing only muddy images.

Query Set -> Gallery Set	Rank1	Rank5	Rank10	mAP
No Mud -> No Mud	0.8852	0.9727	0.9968	0.8809
Mud -> Mud	0.5624	0.7162	0.7349	0.6002
No Mud -> Mud	0.7342	0.8641	0.8543	0.6916
Mud -> No Mud	0.7335	0.8267	0.8333	0.7690

DukeMTMC (Ristani et al., 2016), or Market-1501 (Zheng et al., 2015), we fine-tune the models further on MUDD. The performance of these models is comparable across different source datasets, all substantially improving over off-the-shelf and from scratch approaches.

In summary, pre-training provides significant accuracy gains by overcoming the limited training data through transfer learning. However, gaps to human-level performance remain, motivating techniques tailored to MUDD's extreme conditions. The results reinforce the dataset's unique challenges and domain shift from existing re-id datasets.

5.6 Analysis

Our fine-tuned models demonstrate significant improvements in re-identifying riders compared to off-the-shelf and from-scratch approaches. As seen in Figures 5.4 and 5.7, the model is able to correctly match identities even with mud occlusion if the rider's pose is relatively consistent. This indicates that fine-tuning successfully incorporates invariances to mud while still distinguishing small inter-class differences like gear and outfit.

However, challenges remain under more extreme conditions. In the rest of this section, we analyze several key factors that still cause fine-tuned model failures on MUDD:

Mud occlusion As expected, heavy mud occlusion poses significant challenges. Mud induces high intra-class variation as the amount of mud covering a rider can vary drastically across images. It also causes low inter-class variation since mud occludes distinguishing features like jersey numbers and colors. As shown in Figure 5.6, querying with a muddy image retrieves other muddy images rather than cleaner images of the same identity.

Appearance variation Natural appearance changes of a rider over a race also confuses models. Riders may change gear like goggles or gloves multiple times. Crashes can rip clothing and jerseys. The model must learn to link different levels of mud, gear, and damage of a rider.

Pose variation Complex poses like jumps, crashes, and wheelies are difficult to match, especially combined with mud and appearance variation. As seen in Figure 5.8, a rider doing a wheelie is not matched to more standard riding poses. Even common pose differences like front versus back views are challenging (Figure 5.9).

Low resolution Images with small, distant crops of riders lack fine details for discrimination. Figure 5.11 shows a failure case where the query is low resolution.

Similar outfits In some cases, different riders with very similar gear are confused. This is common as racers supported by the same team will typically purposefully coordinate their appearance. An example is shown in Figure 5.10.

Table 5.2 breaks down the performance of the best fine-tuned re-id model by controlling the query and gallery set for muddy and clean images. We see that by looking only at the clean imagery (i.e. no mud), we get a much better performing model, with gains around 10%. On the other hand, when we evaluate using only the muddy imagery, we see drops in performance of around 20% across the board. Lastly, when a clean or muddy image is used as the query point, and the opposite is used for the gallery set, performance falls between the evaluation using only muddy or only clean images. The accuracy is a bit higher for matching muddy query images to clean gallery images versus the reverse.

We also developed a new data augmentation strategy for dealing with mud occlusion. Inspired by the "splotchy" nature of mud, we introduce the *speckle* data augmentation. This data augmentation in action is shown in Figure 5.12. This technique leads to a 4% improvement in overall Rank-1 accuracy, with the majority of the improvement seen from the mud-occluded images.

In summary, heavy mud occlusion, appearance changes, pose variations, low resolution, and similar outfits remain open challenges. While fine-tuning offers substantial improvements, significant gaps compared to human performance motivates the need for new techniques tailored to these uncontrolled conditions.

5.7 Limitations

While MUDD enables new research into re-id under extreme conditions, our work has several limitations to note:

Labeling Bias Our accelerated labeling methodology leveraging auxiliary information could introduce bias. By searching for images matching the same detected number, we preferentially

sampled identities with more visible numbers. Not only may this over-represent riders with cleaner jerseys and under-represent heavy mud occlusion, but also many riders choose to have very little numbering. The labeling distribution may not fully reflect the underlying data. Models could overfit to the biases of our annotation process. Collecting additional labeled data with different sampling approaches could help quantify and reduce this bias.

Dataset Size MUDD contains 3,906 images across 150 identities. While large for this emerging domain, this remains small compared to widely used re-id datasets. The limited data makes learning robust models difficult. Additional identities and examples would likely improve accuracy, but scaling dataset size is costly in this domain.

Capture Bias All MUDD data was captured during the first half of 2023 across 10 events in the GNCC racing series. This induces bias in the environments, rider identities, and more. Performance may not transfer to other off-road competitions like motocross, supercross, and flat track events. Broader capture diversity could improve model robustness.

Camera IDs MUDD lacks camera ID labels denoting which images came from the same capture device. Camera ID is a useful cue for re-id, enabling models to account for consistent environmental factors and biases per device. However, our dataset combines imagery from 16 different independent photographers at unknown shooting locations.

5.8 Related Work

Person re-identification (re-id) aims to match people across non-overlapping camera views and time horizons (Ye et al., 2021b; Zheng et al., 2016; Zheng et al., 2015; Farenzena et al., 2010). Early re-id methods relied on handcrafted features like color histograms, textures, and local descriptors (Farenzena et al., 2010). With the rise of deep learning, Convolutional Neural Network (CNN) Zhou et al., 2019a and Transformer (He et al., 2021) based approaches now dominate re-id research, spurred by datasets like Market-1501 (Zheng et al., 2015), DukeMTMC-ReID (Ristani et al., 2016), and MSMT17 (Wei et al., 2018).

A few datasets address environmental factors. For example, Xiao et al. (2016) introduce a dataset with a low-resolution challenge set. Occlusions have also been well studied, spearheaded by datasets with high levels of occlusion (Schwartz and Davis, 2009; Wang et al., 2011b; Wang et al., 2016b; Figueira et al., 2015; Xiao et al., 2016). However, these occlusions are unrelated to the heavy mud occlusion in our dataset. The addition of mud drastically complicates re-identification. Furthermore, no prior datasets exhibit such a complex combination of lighting, diversity, motion, and diverse cameras as our off-road racing dataset.

Prior work has focused on the re-identification of motorcycles and bicycles Figueiredo et al., 2021; Li and Liu, 2022; Yuan et al., 2018, however these are restricted to street vehicles in urban settings. A highly related domain is identifying athletes in sports imagery. Penate-Sanchez et al. (2020) release a dataset of ultra-runners competing in a 128km race over the course of a day and a night. While this is more similar to the off-road setting in our dataset, they only have 416 different identities between 5 locations at a single event. Furthermore, there is near zero mud in the dataset. Along similar lines, but in even more controlled and limited settings, are the SoccarNet-ReID (Giancola et al., 2022) and DeepSportRadar-ReID (Van Zandycke et al., 2022) datasets, which contain images from broadcast video of soccer and basketball games respectively.

These datasets have driven research to develop methods to deal with the occlusions common in them. Approaches such as invariant representations (Chen et al., 2019c), metric learning (Yi et al., 2014), semantic attributes (Shi et al., 2015), part-based (Cheng et al., 2016) and pose-aware models Cho and Yoon, 2016, and adversarial learning (Huang et al., 2018) have been proposed to alleviate occlusion problems. Other methods have been developed to handle misalignment, utilizes temporal cues in video (Li et al., 2019), use domain adaptation techniques (Deng et al., 2018), or unsupervised methods (Fan et al., 2018) to reduce label dependencies. Unlike our dataset, these all operate in controlled conditions. Existing models thus fail on our data.

In summary, re-id research has focused on controlled conditions and modest variation. Our dataset introduces real-world challenges absent in existing datasets. Our experiments expose clear gaps between current methods and this application. MUDD provides diverse imagery to spur new techniques for robust re-id under uncontrolled conditions.

5.9 Conclusion

In this work, we introduce MUDD, the first large-scale dataset to benchmark re-identification of motorcycle racers under extreme conditions. MUDD captures challenging factors including heavy mud occlusion, complex poses, variable lighting, and distant perspectives. We propose an accelerated annotation methodology incorporating detected racer numbers to enable efficient high-quality labeling.

Through initial benchmarking experiments, we demonstrate significant gaps between current re-id techniques and the real-world conditions represented in MUDD. Off-the-shelf models fail to generalize to this new domain. Training CNN models like OSNet and ResNet from scratch struggles due to the limited training identities, but fine-tuning pre-trained models on MUDD significantly improves accuracy. Interestingly, models pre-trained on generic ImageNet data prove as effective as re-id-specific pre-training.

However, substantial gaps compared to human performance remain even after fine-tuning. Our analysis reveals open challenges including handling heavy mud occlusion, complex poses,

low-resolution, and similar outfits. These factors induce intra-class variation and inter-class similarity that current models fail to robustly distinguish.

In summary, MUDD exposes clear limitations of existing re-id techniques under uncontrolled conditions. Our work motivates new solutions tailored to the unique challenges of identifying motorcycle racers amidst mud and more. Broader applications such as sports analytics stand to benefit from progress in re-id robustness. MUDD provides diverse, real-world imagery to drive future research towards re-identification in the wild.

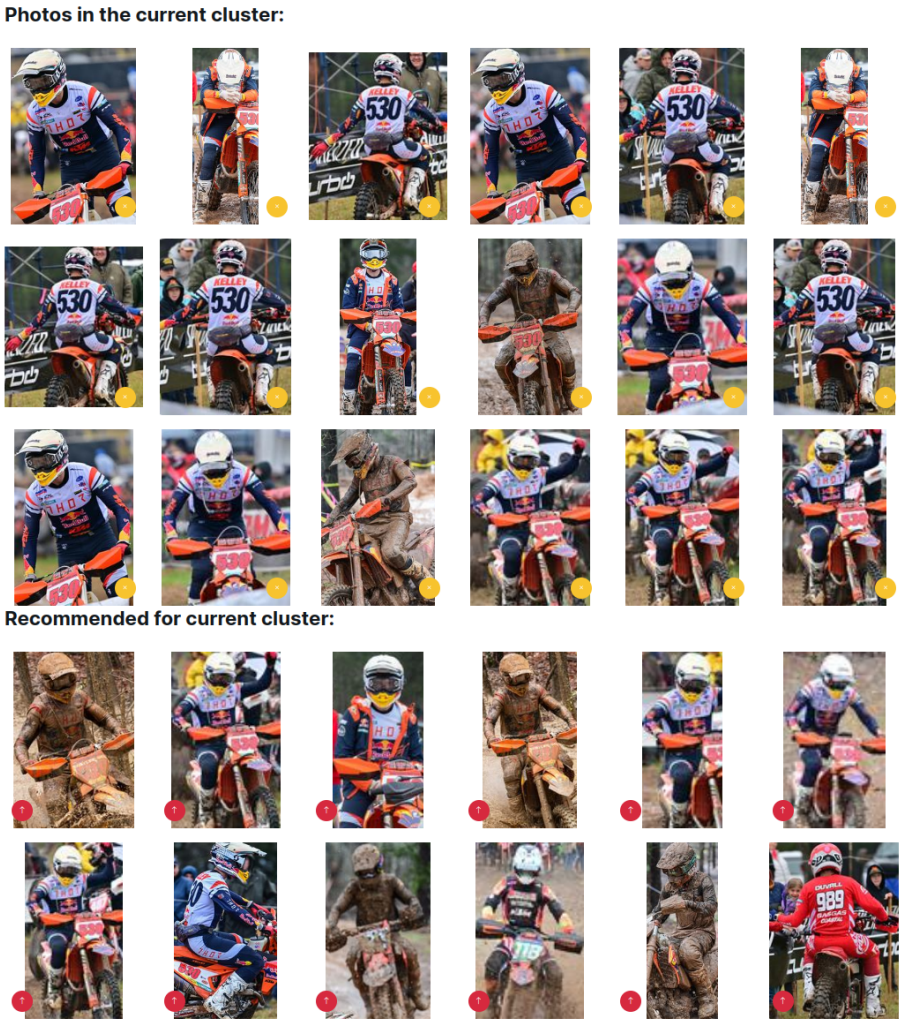


Figure 5.2: Leveraging detected jersey numbers as auxiliary information enables generating higher quality identity clustering proposals for manual verification. This proposed cluster contains both clean and muddy images of the same rider, whereas proposing clusters with off-the-shelf re-id models fail.

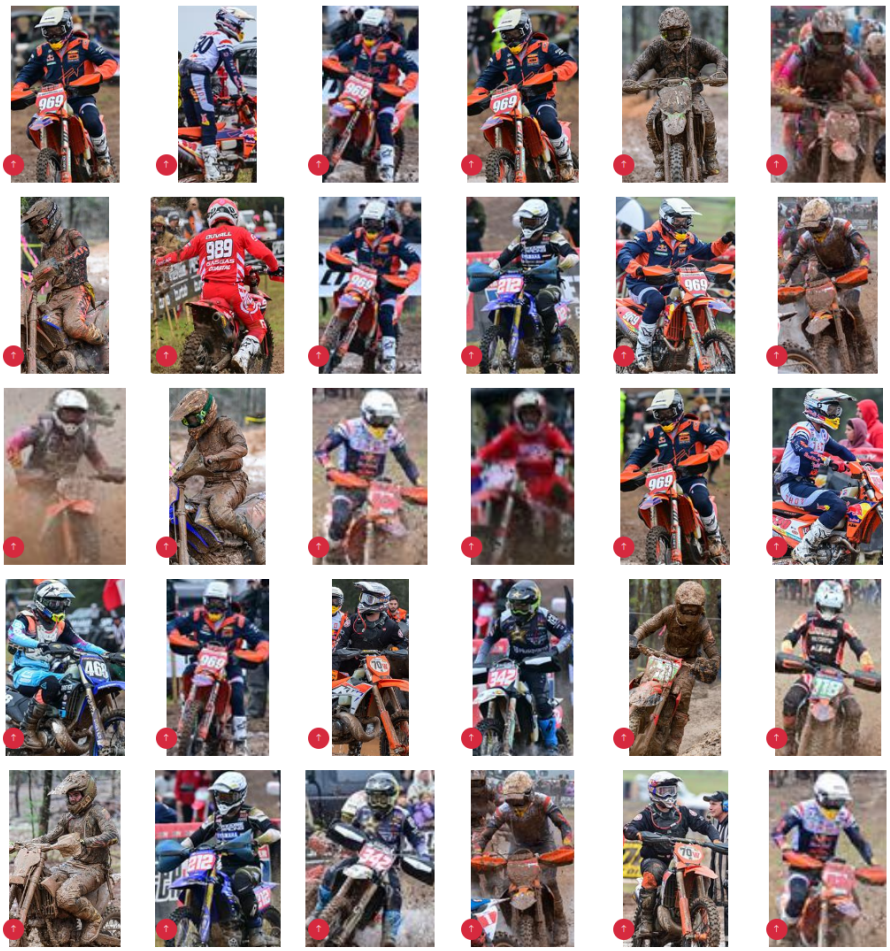


Figure 5.3: Additional proposed results for the same identity cluster as Figure 5.2. Our methodology provides high-quality recommendations to simplify manual verification and labeling.

Figure 5.4: Example of successful re-id by the fine-tuned model under moderate mud occlusion. The 10 top retrievals correctly identify the query rider despite mud, pose, and other variations. Green boundaries signify correct matches and red incorrect.

Figure 5.5: Example of the model correctly matching a clean image of a rider to a muddy image of the same rider when the pose is similar between the query and gallery image. Green boundaries signify correct matches and red incorrect.

Figure 5.6: Failure case with heavy mud occlusion on the query image. Only 1 out of the top 10 results is a correct match, despite over 20 images of the same rider appearing in the gallery set, most of which are clean. Green boundaries signify correct matches and red incorrect.

Figure 5.7: Example of successful re-id by the fine-tuned model under light mud occlusion. All top 10 ranked results correctly match the query rider despite mud, blurring, lighting, pose, and complex backgrounds. Green boundaries signify correct matches and red incorrect.

Figure 5.8: Example of a failure case due to extreme pose variation in the query image. The rider is captured doing a wheelie, leading to incorrect matches despite no mud occlusion. Green boundaries signify correct matches and red incorrect.

Figure 5.9: Failure case due to pose variation between the query and gallery images. The backward-facing query rider is not matched to forward-facing images of the same identity. Green boundaries signify correct matches and red incorrect.

Figure 5.10: Example failure case due to two different riders having very similar jerseys and gear, leading to confusion between their identities. Green boundaries signify correct matches and red incorrect.

Figure 5.11: Failure case due to low resolution of the query image preventing distinguishing details from being visible. The small, distant crop of the rider cannot be matched accurately. Green boundaries signify correct matches and red incorrect.

Figure 5.12: The random speckles data augmentation. Designed to mimic the speckled nature commonly seen from mud.

Part II

Learning to Improve Contrastive Learning

Chapter 6

Risk-Adjusted Mini-Batches

In supervised learning, optimizing the expected loss is common, yet many real-world applications demand attention to tail risks or alignment with human notions of risk. Traditional methods for optimizing these other risks assume full-batch gradient descent, overlooking the complexities of the mini-batch settings critical for deep learning. This study introduces a meta-learning-based approach to derive interpretable mini-batch risk functions, enabling the optimization of diverse risk metrics in a single training process. Our method, Risk-Adjusted Mini-Batches (RAM), demonstrates up to 10% better risk reduction over conventional strategies, significantly enhances model accuracy and robustness against label noise, and can effectively optimize models even when the risk is only implicitly described by a curated subset of data. By meta-learning batch reweightings tailored to specific risk objectives, RAM facilitates more effective optimization of complex risk functions, bridging the gap between theoretical risk sensitivity and practical deep learning applications.

6.1 Introduction

Deep neural networks are typically trained to minimize the expected loss over training data. However, in many real-world settings like healthcare, finance, and transportation, we need to optimize for tail risks and/or align with human notions of risk (Leqi et al., 2019). Consider a cancer detection model. Optimizing for average accuracy fails to capture that missing a tumor is far worse than a false alarm (Patel et al., 2019; Rajpurkar et al., 2020; Tschandl et al., 2020). In lending, focusing only on expected repayment ignores the risk of default (Bussmann et al., 2021; Kruppa et al., 2013).

In such cases, we want to optimize more complex risk objectives like Conditional Value-at-Risk (CVaR) that focuses on high-loss outliers, inverted CVaR (ICVaR) that cares about low-end performance, or human-aligned risks that overweight extreme events (Wong et al., 2022).

Table 6.1: Definitions and interpretations of common risk functions. $F_f(\ell_{f_\theta}(X,Y))$ represents the CDF of $\ell_{f_\theta}(X,Y)$ and $\text{VaR}_\alpha = 100 \times \alpha$-percentile. For more discussion on these risk functions, see Wong et al. (2022).

Risk function	Expression	Interpretation
Expected Value	$\mathbb{E}[\ell_{f_\theta}(X,Y)]$	expected loss
CVaR	$\mathbb{E}[\ell_{f_\theta}(X,Y) \mid \ell_{f_\theta}(X,Y) \geq \text{VaR}_\alpha(\ell_{f_\theta}(X,Y))]$	expectation of losses exceeding the $100 \cdot \alpha$ percentile
Inverted CVaR	$\mathbb{E}[\ell_{f_\theta}(X,Y) \mid \ell_{f_\theta}(X,Y) \leq \text{VaR}_\alpha(\ell_{f_\theta}(X,Y))]$	expectation of losses below the $100 \cdot \alpha$ percentile
Human-aligned	$\mathbb{E}[\ell_{f_\theta}(X,Y) w(F_f(\ell_{f_\theta}(X,Y)))]$	weighting function that overweights extreme losses
Mean-Variance	$\mathbb{E}[\ell_{f_\theta}(X,Y)] + c \cdot \text{Variance}[\ell_{f_\theta}(X,Y)]$	expected loss penalized by its variance
Trimmed Risk	$\mathbb{E}[\ell_{f_\theta}(X,Y) \mid F_f(\ell_{f_\theta}(X,Y)) \in [\alpha, 1-\alpha]]$	ignore extreme losses

However, while optimizing general risk functions is well-studied in the full-batch setting, it remains challenging with mini-batch deep learning. Gradient estimates computed on small batches are often biased for complex risks, causing unstable optimization (Li et al., 2021).

To address this, we propose a meta-learning approach to learn specialized and interpretable reweightings over mini-batches to minimize any differentiable risk function. We refer to this method as Risk-Adjusted Mini-batches (RAM). Given a risk function of interest, we meta-learn a reweighting over quantized batch losses, such that updating the underlying model on the reweighted loss leads to a reduction of the true risk computed on a held-out set.

Experiments on the CIFAR10 and CIFAR100 datasets (Krizhevsky, Hinton, et al., 2009) and six risk functions, discussed in Table 6.1, show RAM reliably optimizes tailored objectives. RAM resulted in up to 10x lower risk, while concurrently improving accuracy by up to 6%. Analysis of the learned risk functions produced by the RAM methodology reveals several interesting behaviors. All learned risk functions exhibit a warm-up period, and then as training progresses, the risk functions specialize according to the problem specifics. While there are similarities between the learned risk functions and the hand-engineered risk functions they are optimizing for, the learned risk functions are surprisingly different.

RAM is also able to effectively learn models among label noise. Even in settings with up to 50% label noise, RAM improves over baseline by up to 10% accuracy (absolute). Even more impressively, no knowledge of the label noise is required, and performance does not degrade over normal learning when no label noise is present. Analysis of the risk functions learned among this label noise reveals behavior mildly characteristic of some hand-engineered, focusing on losses typically characteristic of correctly labeled samples. However, the learned functions are markedly different, especially in their handling of the low-loss samples, the smooth transitions between ignored and emphasized losses, and the occasional multimodality of the weightings.

Overall, we provide a principled approach to risk-sensitive deep learning. By meta-learning batch reweightings tailorable to any differentiable risk metric, we enable effective optimization of diverse risk objectives. In summary, the core contributions are:

- RAM: A novel method to meta-learn Risk-Adjusted Mini-batches tailored to any differentable risk metric.
- Empirical evidence of RAM's efficacy both for risk-sensitive learning and learning among noisy labels.
- Insight into the relationship between mini-batch weightings, model risk, noise labels, and model optimization via analysis of the evolution of the learned risk functions.

6.2 Related Work

Several works have studied optimization for risk metrics beyond expected loss (Duchi and Namkoong, 2021; Duchi et al., 2022), showing benefits for fairness, robustness, and aligning with human risk notions (Leqi et al., 2019). However, optimizing general risk functions with mini-batch, SGD remains challenging.

Prior work developed specialized algorithms for optimizing specific subsets of risk functions. The most general is Tilted Empirical Risk Minimization (Li et al., 2021), which extends empirical risk minimization by adding a "tilt" factor that can be manually tuned to optimize for various types of risk. Curi et al. (2020) propose a learnable sampling strategy for optimizing the CVaR loss. Several other prior works have focused on the optimization of specific risk functions, like the mean-variance tradeoff (Björk et al., 2014) and trimmed loss (Shen and Sanghavi, 2019a; Shen and Sanghavi, 2019b). Instead, we propose a general approach applicable to any differentiable risk metric of interest.

A relevant application is learning with noisy labels. Many works design noise-robust loss functions (Zhang and Sabuncu, 2018; Amid et al., 2019; Lyu and Tsang, 2019), estimate the noise transition matrix (Patrini et al., 2017; Hendrycks et al., 2018; Yao et al., 2020; Yang et al., 2021), or estimate a noise-adapted posterior (Xiao et al., 2015; Han et al., 2018; Yao et al., 2018). We take a more general loss weighting approach requiring no noise modeling.

Importance weighting adapts loss to sample noise levels (Liu and Tao, 2015; Wang et al., 2017; Chang et al., 2017). Jenni and Favaro (2018) meta-learn neural loss correctors, but require clean validation data, MW-Net (Shu et al., 2019) and CMW-Net (Shu et al., 2023) also meta-learn loss weighting functions, reweighting each sample independently. Our work differs by (1) weighting over batch quantiles rather than per-sample losses, (2) focus on risk-sensitive learning, (3) no clean data (i.e. noise-free data) is required. In summary, we introduce a meta-learning framework tailored to risk-sensitive learning. By learning mini-batch reweightings, we can automatically adapt to optimize complex risk functions in deep networks.

Algorithm 1 Meta-Learning Risk Functionals

Require: Initial model parameters θ, initial risk function params ϕ, inner learning rate β, outer learning rate η, validation risk ρ, loss function ℓ, training data $\mathcal{D}$train, validation data $\mathcal{D}$val

while not done **do**
  $\theta' \leftarrow \theta$ {Copy model params}
  for number of inner steps **do**
    Sample batch $B \sim \mathcal{D}$train
    $\hat{y} \leftarrow f\theta'(x_B)$ {Get model output}
    $l_{\text{sorted}} \leftarrow \text{sort}(\ell(\hat{y}, y_B))$ {Sort losses}
    $L_B \leftarrow g_\phi(l_{\text{sorted}})$ {Get minibatch risk}
    $\theta' \leftarrow \theta' - \beta \nabla_{\theta'} L_B$ {Update θ'}
  end for
  Sample batch $V \sim \mathcal{D}$val {Sample validation data}
  $\hat{y} \leftarrow f\theta'(x_V)$ {Get updated model output}
  $l \leftarrow \rho(\ell(\hat{y}, y_V))$ {Compute validation risk}
  $\phi \leftarrow \phi - \eta \nabla_\phi l$ {Update g_ϕ}
  Sample batch $B \sim \mathcal{D}$train {Sample new train batch}
  $\hat{y} \leftarrow f\theta(x_B)$ {Get model output with original θ}
  $\theta \leftarrow \theta - \beta \nabla_\theta g_\phi(\ell(\hat{y}, y_B))$ {Update θ}
end while

6.3 Learning Risk functions

First, we formalize the problem of risk-sensitive learning. Given a training dataset $(x_i, y_i)_{i=1}^n$ comprising n samples drawn from distribution $\mathbb{P}_{\text{train}}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y} \subseteq \mathbb{R}^C$ for C classes, we aim to learn a model f_θ parameterized by θ that minimizes some risk function ρ of the loss ℓ:

$$\theta^* = \arg\min_{\theta \in \Theta} \rho(\ell_{f_\theta}(x, y)), \tag{6.1}$$

where $\ell_{f_\theta}(x, y)$ is the loss incurred by model f_θ on example (x, y), and Θ is a hypothesis class of models.

The standard supervised learning approach sets ρ as the expected value to minimize average loss. However, for many applications like healthcare, finance, and transportation, we may wish to optimize different risk metrics ρ that capture tail risks or align better with human notions of risk. Table 1 summarizes some common choices of ρ.

While optimizing for general risk functions is well-studied in the full-batch setting, it becomes significantly more challenging in the mini-batch setting required for scaling deep learning. With only a small batch $B \sim \mathbb{P}_{\text{train}}$ of training data, the sample mini-batch risk $\hat{\rho}(B)$ will be a biased estimate of the true risk $\rho(\mathbb{P}_{\text{train}})$, making optimization difficult.

Table 6.2: Risk optimization on CIFAR10 for different risk metrics ρ. Our meta-learned approach achieves the lowest risk in nearly all cases, improving over mini-batch baselines by up to 10%.

ρ	Expected Value	batch ρ	batch ρ w/warm-up	Learned
$\mathbb{E}$	**0.268 (0.0047)**	**0.268 (0.0047)**	**0.268 (0.0047)**	**0.278 (0.0047)**
CVaR	2.30 (0.0317)	1.92 (0.0096)	1.77 (0.0098)	**1.72 (0.0156)**
ICVaR	1.46e-6 (1.42e-7)	8.37e-5 (3.26e-5)	1.43e-5 (1.59e-6)	**1.34e-7 (3.23e-8)**
Human	0.595 (0.0061)	**0.576 (0.0076)**	**0.575 (0.0087)**	**0.576 (0.0169)**
Mean-Var	0.342 (0.0087)	**0.329 (0.0109)**	**0.327 (0.0033)**	**0.328 (0.0040)**
Trimmed	0.0480 (0.0013)	**0.0440 (0.0015)**	0.0504 (0.0033)	**0.0460 (0.0026)**

Table 6.3: Accuracy on CIFAR10 when optimizing various risk metrics ρ. Our method maintains competitive generalization despite optimizing complex tailored risk objectives.

ρ	batch ρ	batch ρ w/warm-up	Learned
$\mathbb{E}$	**91.12 (0.290)**	**91.12 (0.290)**	**90.76 (0.470)**
CVaR	68.21 (1.02)	79.14 (0.447)	**85.46 (0.347)**
ICVaR	16.78 (1.48)	89.26 (0.215)	**91.07 (0.0764)**
Human	**90.51 (0.141)**	**90.32 (0.416)**	**90.25 (0.530)**
Mean Var	**91.17 (0.161)**	90.80 (0.204)	90.95 (0.269)
Trimmed	89.05 (0.134)	89.37 (0.172)	**90.52 (0.149)**

To address this, we propose a meta-learning approach to learn a mini-batch reweighting function g_ϕ to provide easier optimization for any risk function of interest. Concretely, g_ϕ outputs a weighted combination of the quantized losses of a mini-batch:

$$L_B = g_\phi(q(\ell_{f_\theta}(x_B, y_B))) \tag{6.2}$$

$q()$ splits them into quantiles, and g_ϕ assigns a learnable weight to each quantile. This allows g_ϕ to weigh the mini-batch losses in a flexible way that reduces the true risk function ρ on a held-out set. Figure 6.1 provides an overview of our approach.

We optimize g_ϕ via bi-level meta-learning. Letting $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$ be the training and validation datasets, the inner loop adapts model parameters θ' on a batch $B \sim \mathcal{D}_{\text{train}}$ by descending the gradient of the mini-batch risk L_B:

$$\theta' = \theta - \beta \nabla_\theta L_B \tag{6.3}$$

The outer loop then computes the true risk $\rho_{\text{val}}(\ell_{f_{\theta'}}(x_{\text{val}}, y_{\text{val}}))$ on the validation data and updates ϕ to reduce it:

$$\phi = \phi - \eta \nabla_\phi \rho_{\text{val}}\Big(\ell_{f_{\theta'}}(x_{\text{val}}, y_{\text{val}})\Big) \tag{6.4}$$

By meta-learning g_ϕ concurrently with model training, we obtain an interpretable risk function specialized for minimizing the true risk metric of interest in a fully online manner.

Table 6.4: Accuracy comparison, given a risk function to minimize ρ on the CIFAR 100 dataset with a Resnet-18 model. The bolded entries represent the highest accuracy models. The reported metrics are averaged over 5 runs, with the standard deviation reported in parentheses.

ρ	batch ρ	batch ρ w/warm-up	Learned
CVaR	61.93 (0.873)	**62.65 (0.412)**	**62.21 (0.890)**
ICVaR	8.158 (2.61)	62.62 (1.85)	**65.97 (0.304)**
Human	**63.55 (0.560)**	63.47 (0.417)	**63.56 (0.269)**
Mean Var	**66.39 (0.280)**	65.38 (0.372)	**66.36 (0.399)**
Trimmed	64.35(0.310)	64.84 (0.196)	**65.29 (0.227)**

Table 6.5: Comparison of the risk of the different methods, given a risk function to minimize ρ on the CIFAR 100 dataset with a Resnet-18 model. The lowest-risk entries are bolded. The reported metrics are averaged over 5 runs, with the standard deviation reported in parentheses.

ρ	$\mathbb{E}$	batch ρ	batch ρ w/warm-up	Learned
CVaR	5.204 (0.071)	5.285 (0.0614)	4.126 (0.0118)	**4.077 (0.0138)**
ICVaR	1.22e-3 (2.75e-5)	1.347e-3 (2.30e-4)	0.6337 (0.2602)	**1.576e-4 (8.21e-5)**
Human	1.895 (0.010)	1.841 (0.010)	**1.746 (0.00894)**	**1.746 (0.00910)**
Mean Var	1.458 (0.023)	1.455 (0.0203)	**1.423 (0.0122)**	**1.426 (0.0156)**
Trimmed	**0.8239 (0.010)**	0.8688 (0.00873)	0.9299 (0.0148)	**0.8292 (0.0192)**

6.4 Experimental Evaluation

We conduct experiments to evaluate RAM for optimizing risk functions and analyze their behavior among label noise. All code, data, and experiments can be found at https://drive.google.com/file/d/1ssMJe0ADy6vjNbJ_GdmNvu8JQHfipb6t/view?usp=drive_link.

6.4.1 Experimental Setup

We evaluate our proposed RAM on image classification using CIFAR10 and CIFAR100 datasets (Krizhevsky, Hinton, et al., 2009). For the model architecture, we use ResNet-18 (He et al., 2016) due to its widespread adoption. For hyperparameter selection, we performed a grid search as detailed in Table 6.6. A 1-cycle policy learning rate scheduler was used (Smith and Topin, 2019). The searches were run over 20,000 steps, consuming 576 GPU hours on 8 V100 NVIDIA GPUs, evaluating on a separate help-out hyper-validation set. We implement the learned risk function g_ϕ as a linear layer without biases followed by a softmax. This produces a convex combination of the input loss quantiles. We bucket the batch losses into $Q = 100$ quantiles sorted from highest

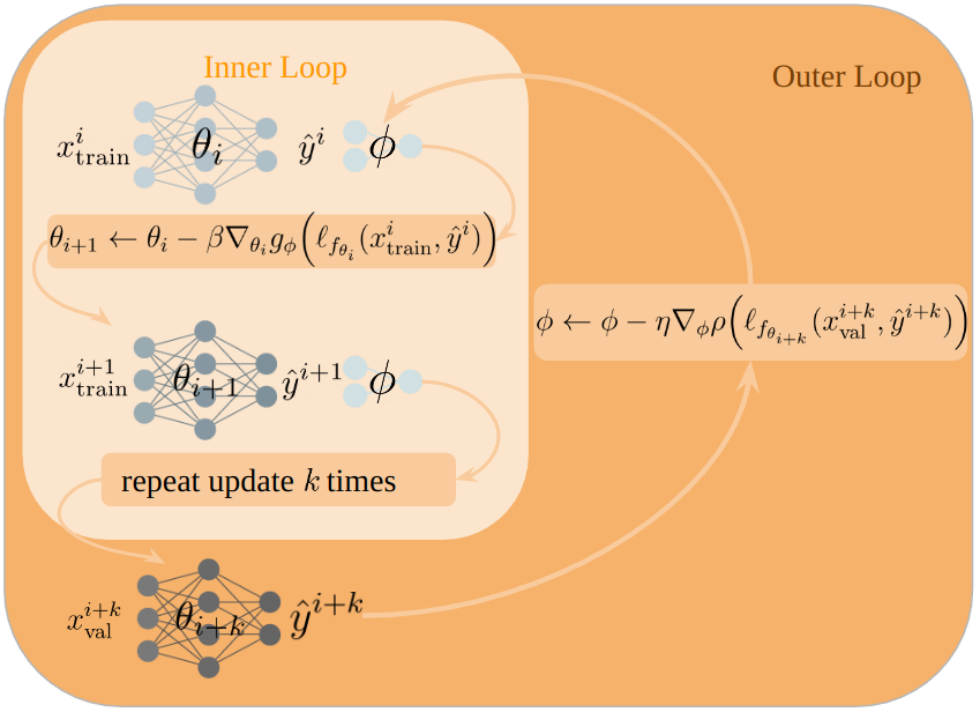
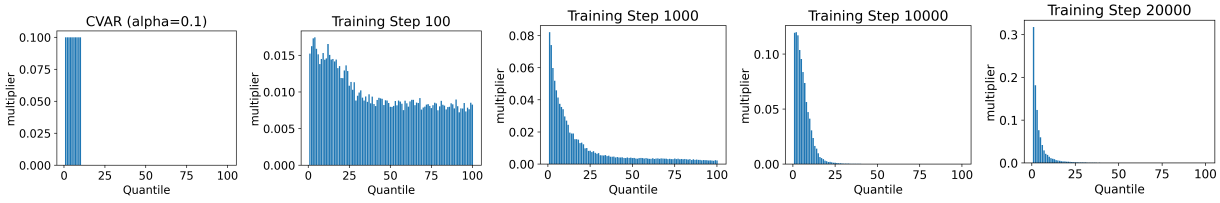


Figure 6.1: Overview of our meta-learning approach. Given a model f_θ and inner/outer learning rates β/η, we meta-learn a mini-batch risk function g_ϕ that outputs a weighted combination of sorted loss quantiles. g_ϕ is trained to minimize the validation risk ρ of interest. This provides a way to optimize complex risk objectives and adapt to distribution shifts.

to lowest loss.

For evaluation, we report metrics on the test set after retraining on the combined train and validation sets, using the best hyperparameters from the validation set. We ran each experiment 5 times with different random seeds and report the average and standard deviation.

6.4.2 Risk Optimization Experiments

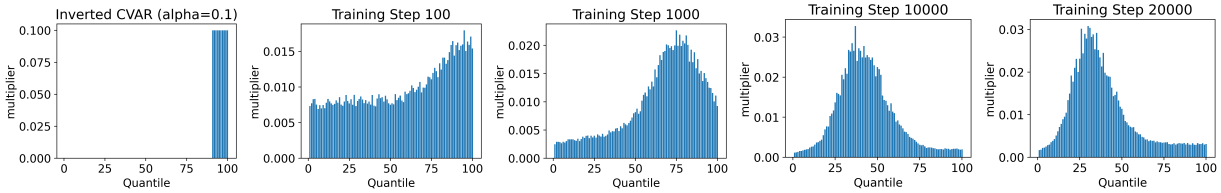


(a) CVaR (b) t=100 (c) t=1000 (d) t=10000 (e) t=20000

Figure 6.2: Learned mini-batch risk functions when optimizing CVaR on CIFAR10. The risk weightings exhibit a warm-up period then concentrate on high-loss samples relevant for CVaR.

Table 6.6: The search ranges for hyperparameter optimization, along with the best-performing hyperparameters (Final Value).

Parameter	Search Range	Final Value
momentum	$[0, 0.99]$	0.9
weight decay	$[0, 1e-2]$	5e-4
batch size	$[20, 200]$	100
min learning rate	$[1e-7, 0.01]$	5e-6
max learning rate	$[0.001, 0.5]]$	0.1
starting learning rate	$U(0.0001, 0.01)$	0.005
inner steps	$[1, 20]$	5
inner learning rate	$[0.0001, 0.1]$	0.001



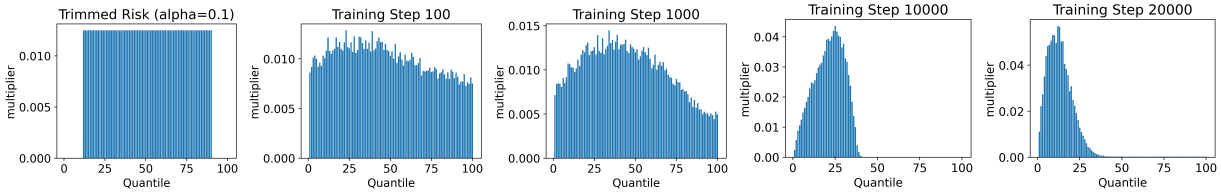
(a) ICVaR (b) t=100 (c) t=1000 (d) t=10000 (e) t=20000

Figure 6.3: Learned mini-batch risk functions when optimizing ICVaR on CIFAR10. The risk weightings exhibit a warm-up period and then concentrate on middle-loss samples.

We evaluate how effective RAM is at directly optimizing the commonly used risk metrics detailed in Table 6.1. For each risk metric ρ, we compare four training approaches:

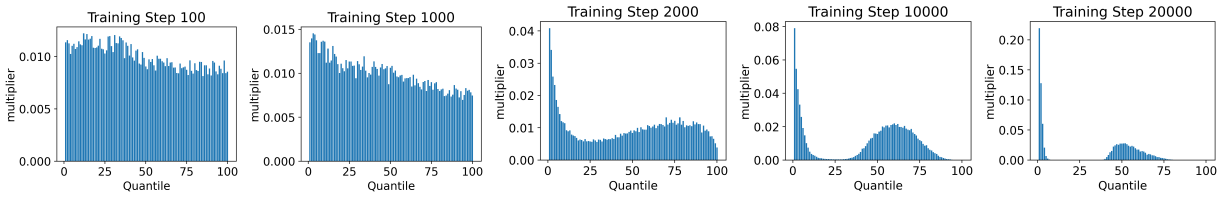
- **Standard training**: minimize the expected mini-batch loss
- **Sample** ρ: Replace the mini-batch loss with a sample estimate of ρ computed on the batch
- **Warm start**: Warm start with standard training, then fine-tune with sample ρ
- **Ours**: Meta-learn the mini-batch risk function g_ϕ to minimize the validation ρ

Table 6.2 compares these training approaches on the various risk functions on the CIFAR10



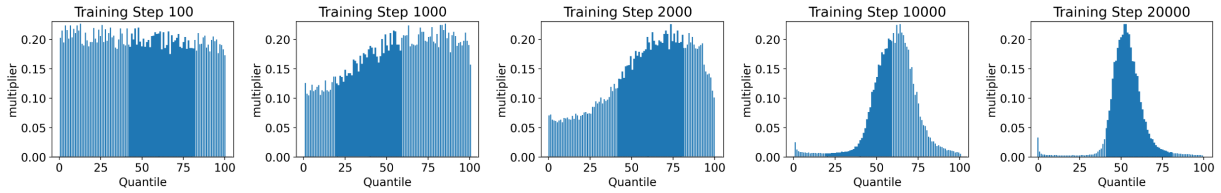
(a) Trimmed (b) t=100 (c) t=1000 (d) t=10000 (e) t=20000

Figure 6.4: Learned mini-batch risk functions when optimizing trimmed risk on CIFAR10. The risk weightings exhibit a warm-up period then avoid the most extreme high/low losses.



(a) t=100 (b) t=1000 (c) t=2000 (d) t=10000 (e) t=20000

Figure 6.5: Learned mini-batch risk functions on CIFAR10 with 50% label noise. The risk weightings exhibit a warm-up period then focus on clean-labeled samples.



(a) t=100 (b) t=1000 (c) t=2000 (d) t=10000 (e) t=20000

Figure 6.6: Learned mini-batch risk functions on CIFAR10 with 50% label noise, but a clean validation set. The risk weightings exhibit a warm-up period then focus on clean-labeled samples.

dataset. In every case, RAM produces models with (or matching) the lowest risk. On the human, mean-variance, and trimmed risk functions, simply applying them at the batch level is effective, resulting in models with the same risk levels as RAM. Likely, this is due to the less biased estimates they provide when applied to data subsets. However, we see that the CVaR and ICVaR risk functions applied at the batch level result in suboptimal, and therefore more risky, models. RAM improves over CVaR baselines by 3% and reduces the ICVaR to under 1% (100x relative improvement) of its original value. Table 6.5 provides the same results but on the CIFAR-100 dataset, consisting of 10x more classes than CIFAR10. Again, we come to the same conclusions. RAM remains the lowest risk method for each of the risk functions, providing the most drastic improvement over baselines for ICVaR.

Another important consideration when training risk-sensitive models is to measure how other performance measures are affected. Occasionally, the risk is the end-all-be-all metric to optimize, but more often than not, a balance between risk and accuracy must be struck. Table 6.3 shows the accuracy corresponding to all of the settings just discussed in Table 6.2, on the CIFAR10 dataset. In all cases except one, RAM yields the highest accuracy models with increases in accuracy of up to 6%. The one exception is optimizing with the mean-variance risk function applied at each batch. More research is needed to understand why this is the case. Again, as shown in Table 6.4 these results are mimicked for the CIFAR100 dataset. RAM not only yields the lowest risk models but also the most accurate.

A side-effect of the simple network used for the learned risk function g_ϕ is that by looking at the weights we can get a sense of what is happening. The losses are sorted before they are

given to the learned risk function, so plotting ϕ gives a graphical representation of how the high and low losses are treated within each batch. Figures 6.2-6.4 show exactly this, in addition to the ρ of interest depicted in the same manner as the left-most plot.

Figure 6.2 details the learned mini-batch risk function (parameters ϕ) at different points in training when optimizing the CVaR on CIFAR10. phi is initialized randomly, and despite a high enough learning rate to vary drastically in few steps, phi remains relatively uniform early on in training, as highlighted by Figure 6.2b. As training progresses, ϕ begins looking more and more like the underlying CVaR risk function, but with one primary difference: the CVAR uniformly weights the highest losses of interest, but ϕ weights the highest loss samples much more than the still-high-but-not-highest samples. Intuitively, this makes sense because of uncertainty. It is impossible to tell the exact cutoff within a randomly drawn minibatch, so a reasonable strategy is to weight the highest losses very high and then taper off the loss centered around the best approximation of where the losses no longer matter.

The learned ϕ for ICVaR differs greatly. Instead of moving closer and closer to the ICVaR risk function, ϕ overweight the middle quantized losses, and underweighting the highest and lowest. The most likely explanation for this is that focusing only on the lowest losses greatly prohibits model training and amplifies overfitting. Therefore, and perhaps slightly counter-intuitively, to minimize the lowest losses, signals from the entire spectrum of quantized losses must be used to build an inherently better model.

Finally, looking at the phi generated when optimizing models for the trimmed loss, Figure 6.4 highlights a similar case as CVaR, where ϕ starts off mostly uniform and then transitions towards the trimmed risk function. However, Figure 6.4e shows late in training, where still see behavior representative of ignoring the highest and lowest losses, but in a very different range than the underlying trimmed risk function. There are likely some artifacts in these learned risk functions. In an ideal world, the number of inner steps used would be the number of inner steps needed for a model to converge. Of course, this is impractical both in terms of time and the computational requirements needed to maintain such a computation graph (as you'd need to backpropagate through the entire thing). To make this feasible, we use a small number of inner steps, 5 in the cases of these figures. Therefore, the risk functions being learned are optimized to maximize performance in the next 5 steps. Nevertheless, even with this limitation, we can train performant models. Furthermore, one notable finding is the surprising differences between the learned risk functions ϕ and the underlying risk function ρ. These results indicate that a large difference would still appear, even without the inner step limitation.

Analysis of the learned risk functions resulting from RAM yields several interesting behaviors. All learned risk functions exhibit a warm-up period early in training where all samples are weighted roughly equally. Initially, the model is random so relying solely on extreme samples would not be as informative. As training progresses, the risk reweightings specialize based on the validation risk objective. While there are similarities between ϕ and ρ, ϕ differs markedly from ρ in every case. ϕ shows training phases corresponding to different stages of model optimization, transitioning smoothly between them by the gradual evolution of the risk reweight-

ings. ϕ can change much more quickly than the model parameters θ, yet we see smooth and slow changes regardless.

6.4.3 Learning Risk Functions for Label Noise

We now examine RAM for handling label noise, a common form of distribution shift. While we do not expect to beat specialized state-of-the-art techniques, this allows us to gain further insights into the learned mini-batch risk functions. We introduce varying levels of random label noise to CIFAR10, randomly flipping labels to incorrect values. We compare against two baselines:

- **Standard training**: Mini-batch expected loss
- **Oracle**: Optimizes the ICVaR risk with α set to the true noise rate. Represents the theoretically best possible performance.

The oracle uses the noise rate which is unknown in practice, whereas RAM automatically adjusts to the noise without knowledge of the rate. Table 6.7 shows results utilizing a small clean validation set (2% of train data). RAM substantially improves robustness over standard training, approaching the oracle despite not utilizing the noise rate. Without any noise, our method matches standard training accuracy. At 50% noise, it achieves 68.75% accuracy, compared to just 60.18% for standard training. Interestingly, Table 6.8 shows our method is effective even without any clean validation data, suffering virtually no loss in performance. RAM can improve robustness even when the label noise is unknown and present in the validation set. Real-world datasets typically contain some level of noise. RAM provides a learning method that will not hurt performance vs normal learning, and can simultaneously offer large benefits if label noise is present even without knowledge of it.

Following our analysis of the learned parameters ϕ in the risk-sensitive setting, Figures 6.5 and 6.6 gives the same depiction but during training among noisy labels. Figure 6.6 details ϕ as it evolves during training, when the validation set contains *no* noise. Initially, the weighting starts roughly uniformly but quickly progresses to focusing on the lower 50% of losses, which is the theoretically optimal approach. Later in training, we see that the lowest 25% of losses, and the highest 50% of losses are largely ignored, while all of the focus is placed roughly in the 50% to 75% of losses range. Without the presence of a noise-free validation set (i.e. Table 6.5), again the learned weighting learns to focus more on the lowest 50% of losses after an initial warm-up period. Then later in training, focus is placed roughly on the 40'th to the 75'th percentile. Intriguingly, the highest few losses are always weighted highly. This could be due to the noisy signal in the validation, falsely tricking the learned risk function to optimize the high-loss samples. However, more research is needed to understand this phenomenon.

Table 6.7: Test accuracy on CIFAR10 under different label noise rates, given a small clean validation set. Our method improves robustness to noise without requiring the noise rate.

Random Labels	Expected Value	Learned	Oracle
0%	**91.25 (0.241)**	**91.08 (0.277)**	**91.25 (0.241)**
5%	80.98 (0.161)	**82.65 (0.549)**	85.13 (0.201)
20%	71.75 (0.909)	**79.02 (0.4807)**	82.92 (0.291)
50%	60.18 (2.94)	**68.75 (1.42)**	79.14 (0.371)

Table 6.8: Test accuracy on CIFAR10 under different label noise rates, without using any clean validation data. Our method remains effective at handling noise.

Random Labels	Expected Value	Learned	Oracle
0%	**91.25 (0.241)**	**91.08 (0.277)**	**91.25 (0.241)**
5%	80.62 (0.498)	**81.71 (0.289)**	84.07 (0.189)
20%	68.46 (0.394)	**77.85 (0.557)**	82.31 (0.376)
50%	60.72 (2.46)	**69.54 (1.23)**	76.86 (0.311)

6.5 Learning Mining Functions

A mining function is a risk function but applied in a contrastive learning context. Instead of weighting each sample in a mini-batch to better align a model with a notion of risk, contrastive learning must contend with optimizing from pairs or triplets of data. Given a batch of samples, all possible pairs or triplets are formed, and the loss for each is calculated. Traditionally, the losses were summed or averaged, and used to update the underlying model. However, it was found that this leads to an abundance of easy (i.e. low loss) pairs/triplets, but few hard (i.e. high loss) pairs/triplets. So instead of summing or averaging, simply selecting the *hardest negative* example for each pair/triplet resulted in much more efficient and effective learning, and ultimately better-performing models. This is referred to as *hard-negative mining*. Note, however, that there is still an averaging operation that takes place, even in the case of hard-negative mining. With hard-negative mining, each sample in a batch is paired with the hardest negative instead of all possible negative samples, but each sample in the batch still has a loss. So the mean operation is used to reduce the loss from each pair/triplet down to a single loss for model updating.

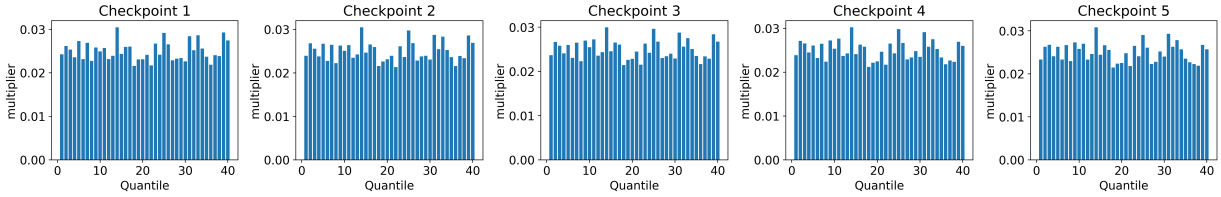
But hard-negative mining is not the end-all-be-all. There are cases where it is *much* worse than using the normal summing/averaging approach. Some intuition exists between hard negative mining, semi-hard negative mining, and other hand-engineered mining functions, but the mining function becomes another hyperparameter to search over. Furthermore, while we have a few of these hand-engineered mining functions, it is unclear if any better ones exist, or what

Table 6.9: Rank-1 (R1) accuracy and mean average precision (mAP) for models learned with hand-engineered hard-negative mining function versus learning a mining function.

Dataset	Method	R1	mAP
Market1501	Hard Negative Mining	94.8	84.9
	Learned Mining Function	94.9	84.8
MUDD	Hard Negative Mining	0.793	0.827
	Learned Mining Function	0.833	0.860

they may be. In this section, we use RAM to directly optimize for useful mining functions. For experimentation in this area, we use the Market-1501 (Zheng et al., 2015) and MUDD (Tyo et al., 2023) person re-identification datasets. The results on both of these datasets, both when using the normal hard-negative mining and when learning a mining function are shown in Figure 6.9.

First, we explore using RAM *alongside* hard negative mining. Instead of using the mean operation as the final step in generating the batch loss, we use RAM just as it is used throughout the rest of this paper. Interstingly, RAM provides no benefit over the normal mean operation, only matching its performance in the best case. Figure 6.7 depicts the risk functions learned in this case, as you can see, it just noisily oscillates around the simple mean operator.

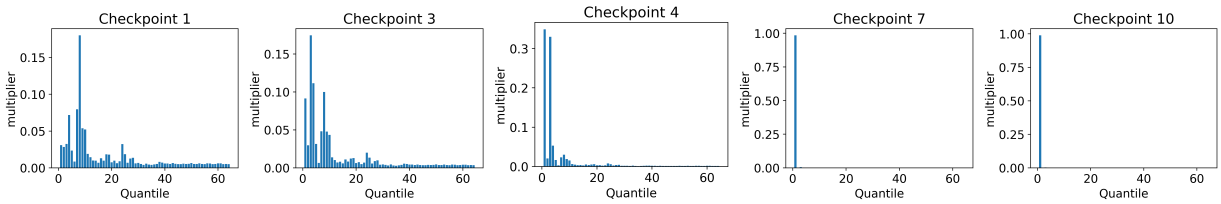


(a) t=100 (b) t=1000 (c) t=2000 (d) t=10000 (e) t=20000

Figure 6.7: The *batch reduction* function optimized to maximize the rank-1 accuracy on the Market-1501 dataset, given a batch of triplets constructed from the hardest negative for each anchor.

Second, we explore using RAM in a way more representative of a true mining function – all possible pairs/triplets are formed, and then RAM is used to reduce the loss for each of them into a single loss for the sample. On the market-1501 dataset, we see that RAM starts with an emphasis on the high-loss samples, but quickly learns the exact hard-negative mining strategy, as shown in Figure 6.8. This is an interesting finding, but it was on the Market1501 dataset, which is a very clean dataset. To explore how the learned mining function shapes up in a less ideal setting, the MUDD dataset, which has a small amount of label noise. However, the same phenomenon is observed.

To further explore the power of learned mining functions, we experiment by adding different amounts of label noise to the Market1501 dataset. Table 6.10 details the performance of the various settings, while Figure 6.9 shows the characteristics of the learned mining function



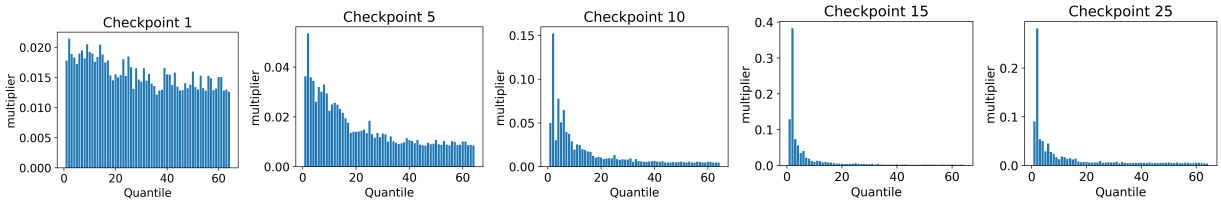
(a) t=100 (b) t=1000 (c) t=2000 (d) t=10000 (e) t=20000

Figure 6.8: The *mining function* optimized to maximize the rank-1 accuracy on the Market-1501 dataset, given all possible valid triplets for each anchor point. The resulting loss for each anchor in the batch was then averaged to form the final loss.

Table 6.10: Rank-1 (R1) accuracy and mean average precision (mAP) for models learned with hand-engineered hard-negative mining function versus learning a mining function on the Market1501 dataset with varying levels of label noise.

Noise	Method	R1	mAP
20%	Hard Negative Mining	67.3	69.9
	Learned Mining Function	74.2	77.3
40%	Hard Negative Mining	0.007	0.010
	Learned Mining Function	0.433	0.452

among 20% label noise, and Figure 6.10 shows the same among 40% label noise.

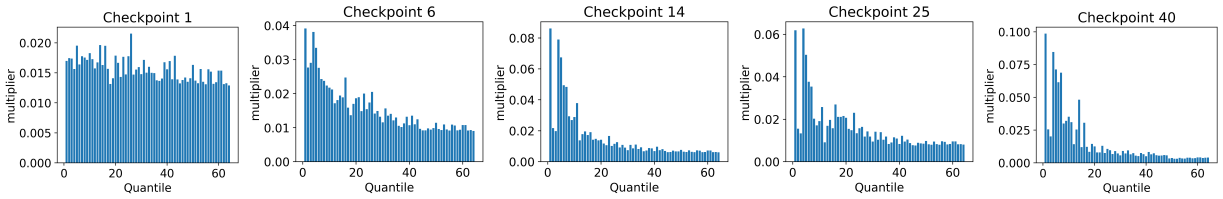


(a) t=100 (b) t=1000 (c) t=2000 (d) t=10000 (e) t=20000

Figure 6.9: The *mining function* optimized to maximize the rank-1 accuracy on the Market-1501 dataset among 20% label noise, given all possible valid triplets for each anchor point. The resulting loss for each anchor in the batch was then averaged to form the final loss.

6.6 Conclusion

In this work, we introduced Risk-Adjusted Mini-Batches (RAM), a meta-learning-based approach for learning specialized and interpretable mini-batch risk functions for risk-sensitive learning. RAM uses bi-level learning to optimize for a mini-batch reweighting over quantized mini-batch losses, where the inner objective is to minimize the reweighted loss, and the outer



(a) t=100 (b) t=1000 (c) t=2000 (d) t=10000 (e) t=20000

Figure 6.10: The *mining function* optimized to maximize the rank-1 accuracy on the Market-1501 dataset among 40% label noise, given all possible valid triplets for each anchor point. The resulting loss for each anchor in the batch was then averaged to form the final loss.

objective is to minimize any hand-engineered risk function of interest. RAM produced models with 10x lower risk than baselines, on multiple datasets, when optimizing for the Inverted Conditional Value-at-Risk (ICVaR). While the improvement on the other risk functions measured was not as dramatic (typically around 3-5%), these improvements occur concurrently with up to a 6% improvement in accuracy. Analysis of the learned risk functions produced by the RAM methodology reveals several interesting behaviors. All learned risk functions exhibit a warm-up period, and then as training progresses, the risk functions specialize according to the problem specifics. While there are similarities between the learned risk functions and the hand-engineered risk functions they are optimizing for, the learned risk functions are surprisingly different.

RAM is also able to effectively learn models among label noise. Even in settings with up to 50% label noise, RAM improves over baseline by up to 10% accuracy. Even more impressively, no knowledge of the label noise is required, and performance does not degrade over normal learning when no label noise is present. Analysis of the risk functions learned among this label noise reveals behavior mildly characteristic of some hand-engineered, focusing on losses typically characteristic of correctly labeled samples. However, the learned functions are markedly different, especially in their handling of the low-loss samples, the smooth transitions between ignored and emphasized losses, and the occasional multimodality of the weightings.

Overall, we presented a principled approach for risk-sensitive deep learning. By meta-learning batch reweightings tailored to a risk metric of interest, we enable the optimization of diverse objectives capturing tail risks, human notions of risk, and/or much more. We look forward to future work investigating the effectiveness of learning transferrable mini-batch reweightings via RAM to specific risk-sensitive settings.

Chapter 7

Contrastive Multiple Instance Learning for Weakly Supervised Person ReID

The acquisition of large-scale, precisely labeled datasets for person re-identification (ReID) poses a significant challenge. Weakly supervised ReID has begun to address this issue, although its performance lags behind fully supervised methods. In response, we introduce Contrastive Multiple Instance Learning (CMIL), a novel framework tailored for more effective weakly supervised ReID. CMIL distinguishes itself by requiring only a single model and no pseudo labels, while leveraging contrastive losses – a technique that has significantly enhanced traditional ReID performance yet is absent in all prior MIL-based approaches. Through extensive experiments and analysis across three datasets, CMIL not only matches state-of-the-art performance on the large-scale SYSU-30k dataset with fewer assumptions but also consistently outperforms all baselines on the WL-market1501 and Weakly Labeled MUddy racer re-iDentification dataset (WL-MUDD) datasets. We introduce and release the WL-MUDD dataset, an extension of the MUDD dataset featuring naturally occurring weak labels from the real-world application at PerformancePhoto.co.

7.1 Introduction

Accurate data labeling is a critical part of any machine-learning system, but is often prohibitively expensive, especially for person re-identification (ReID). In most classification problems, the classes are easily human-recognizable, allowing annotators to quickly recognize and label the class of a data point. In ReID however, the data points can consist of millions of individuals, none of which are known to the annotators. In this case, generating accurate labels is extremely difficult and time-consuming. An alternative approach is to use weakly supervised learning (WSL) methods that can effectively leverage lower-quality data labeling, which is of-

ten available in larger amounts at meager cost (Liu et al., 2023; Wang et al., 2021b; Meng et al., 2019; Zhao et al., 2021; Wang et al., 2020a).

WSL has achieved impressive results on benchmark datasets, but performance still lags that of the standard, fully-supervised, setting. Given that the reid task of identifying images of the same person is inherently contrastive with respect to identities, it seems possible that we could leverage techniques from contrastive learning to improve WSL further. Contrastive learning is a specific subset of supervised learning where models are optimized on pairs (or triplets) of inputs to determine if the inputs originate from the same class or not. This slight reframing has major benefits both in terms of model performance and generalization (Garg et al., 2023; Hermans et al., 2017), and in terms of computational efficiency in downstream applications. Specifically for ReID, common applications include facial recognition, person search, and image retrieval. In each of these settings, the number of downstream classes (identities) is typically unknown, a setting where contrastive models excel. However, traditional algorithms in contrastive learning depend on accurate labels.

Weak labels for ReID can be gathered in several ways. One example, as provided by Guillaumin et al. (2010), is to gather images of people based on an online search. The resulting dataset is bags of images that all contain the same person, but the images would also be extremely noisy, containing many other people in each photo. Another example of this type of weak labels is to observe event photo purchases - someone purchasing photos of a racer after a marathon likely purchase photos that all contain a common single person. As part of this work, we introduce and release the Weakly Labeled MUddy racer re-iDentification dataset (WL-MUDD) dataset, which is a dataset labeled in this exact manner from the motorcycle racing event photo website PerformancePhoto.co.

The dominant methods for weak ReID rely on pseudo-labeling (Liu et al., 2023; Zhao et al., 2021; Zheng et al., 2021; Wang et al., 2020a), which is an iterative process of predicting new labels for the weakly labeled data in an attempt to build better models. Other approaches include graph-based methods (Wang et al., 2021b; Meng et al., 2021), Multiple Instance Learning (MIL) (Huang et al., 2017; Wu et al., 2015; Sudharshan et al., 2019), or transferring an unsupervised model (i.e. trained without labels). The unsupervised methods have made significant progress recently, but still fall short of methods that can leverage labeling (Wang et al., 2020a). Within WSL, pseudo-labeling approaches typically outperform those of noisy learning and MIL. However, the existing MIL formulations restrict the use of contrastive methodologies.

In this work, we introduce Contrastive Multiple Instance Learning to enable contrastive learning among weakly labeled bags of images. Contrastive learning is typically interpreted as decreasing the distance between the representation of two images of the same identity (or class), and increasing the distance between the representation of two images of different identities. However, in weakly supervised learning, the labels are not that granular. Pseudo-labeling methods get around this by trusting that the labels are granular enough, and then updates the labels as training progresses, but this is prone to errors. Especially in settings where the intra-identity variability is extremely high, and the inter-identity variability is low, which is the exact

case for our WL-MUDD dataset. Instead of a label refinement approach, we focus on the MIL formulation, and to enable contrastive techniques, we formulate the contrastive learning problem as decreasing the distance between two *bag* representations that have the same label, and increase the distance between two *bag* representations with a different label.

This shift in perspective, of optimizing for bag representations instead of representations of a single image within that bag, is not obvious, mainly because at test time, the goal is still to produce a high-performing ReID model: one that can take a single image and produce a high-quality embedding for it. To this end, CMIL includes two processes to help in this regard. The first is that each image in each bag is independently embedded into a representation using a feature extraction network, resulting in a bag of image features. Then, the bag of image features is passed through an accumulation network to generate a bag representation. Second, we experiment with an *alignment loss*, to encourage our model to learn image and bag representations that are similar.

The feature extraction network is chosen to be a standard ReID model, specifically ResNet-50 (He et al., 2016). The accumulation network must be permutation invariant, and is therefore chosen to be a set transformer, although we do provide ablation studies with the simpler choices of the average, max, and sum operators. Surprisingly, we find that even without the alignment loss, optimizing for high-quality bag representations implicitly leads to high-quality image representations.

We evaluate CMIL against a state-of-the-art weakly supervised learning method (Ye et al., 2021a) and a prior MIL method (Meng et al., 2019) on the weakly-labeled Market1501 (WL-Market1501) dataset, and WL-MUDD datasets. The WL-Market1501 dataset is the widely used Market1501 dataset (Zheng et al., 2015) but with noise added to mimic the weakly labeled setting. Then, we compare CMIL to the state-of-the-art on the large-scale SYSU-30k weakly labeled ReID dataset, containing nearly 30 million images and over 30 thousand identities. We find that on both WL-Market1501 and WL-MUDD, CMIL consistently achieved the best rank-1, rank-5, rank-10 accuracy, and mean accuracy precision. On the SYSU-30k dataset, CMIL matched the state-of-the-art while requiring fewer modeling assumptions. Lastly, our ablation studies reveal the surprising effectiveness of average pooling for image aggregation, along with the surprisingly different instance and bag representations even of the best-performing models.

The contributions of this work are threefold:

- The introduction and release of WL-MUDD, a real-world dataset of motorcycle racers with naturally weak labels from PerformancePhoto.co.
- We introduce CMIL, a novel framework for re-identification from weakly labeled group images.
- Experimental evidence of the efficacy of CMIL and an analysis highlighting the surprising differences between image and bag representations.

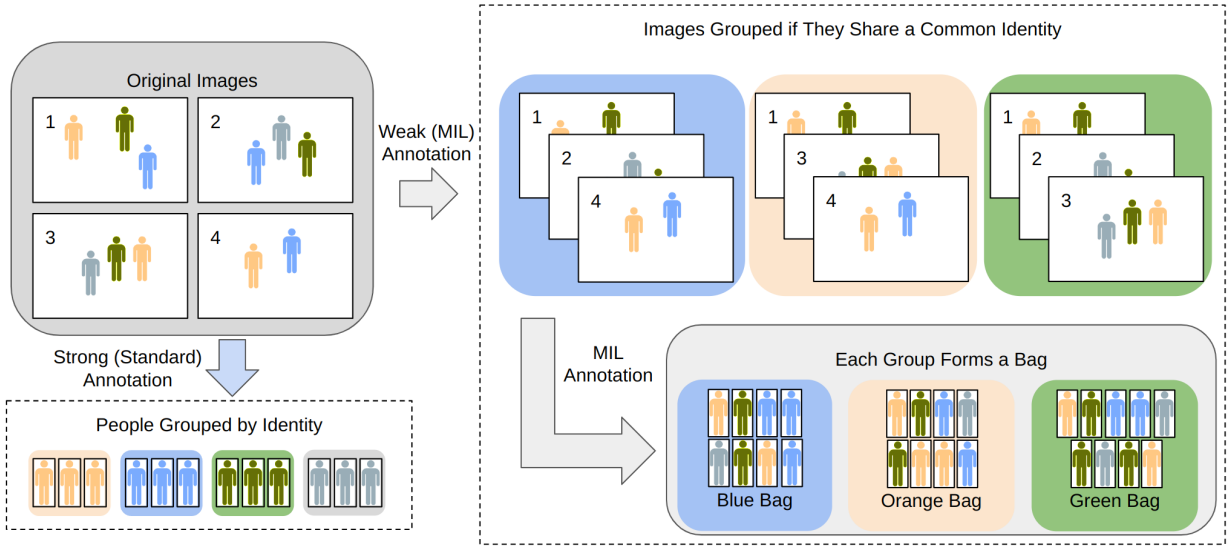


Figure 7.1: The annotation process for strong and weak ReID. The strong annotations group each crop into a bag based on their identity, whereas the weak annotation groups all images based on a shared identity, and then all crops from the grouped images become a bag.

7.2 Datasets and Problem Setup

In this section, we formally introduce the weakly supervised ReID setting, as well as a new weakly supervised ReID dataset. The dataset is available at https://drive.google.com/file/d/1rjMbWB6m-apHF3Wg_cfqc8QqKgQ21AsT/view?usp=drive_link.

7.2.1 Weakly Supervised Re-Identification

The problem we address in this paper is re-identification from weakly labeled group images. We assume that we are given a dataset of images $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ where each image I_j contains one or more people that we are interested in identifying. Let $X_j = \{x_1^j, x_2^j, \ldots, x_{M_j}^j\}$ denote the set of M_j *crops* containing each person extracted from image I_j. We refer to each specific person in an image I_j as $x_i^j \in X_j$.

However, unlike conventional re-identification datasets, we only have weak labels for each image. These labels merely indicate the presence of a shared identity within each group, but not the specific identity of each individual instance within the group. This means that we have access to a set of *bags* $\mathcal{B} = \{B_1, B_2, \ldots, B_K\}$ where each bag $B_k = \{I_{k_1}, I_{k_2}, \ldots, I_{k_{|B_k|}}\}$ contains images that share a common identity. Importantly, the individual instances within each group are not labeled with their specific identities. Instead, the bag is labeled with only a single identity. The key challenge in this setting is to learn a model that can effectively discriminate between different identities despite only having access to these weak bag-level labels. During

inference time the bag-level labels are not the label of interest. Instead, we want a standard ReID model at inference time, meaning that we need to be able to predict the identity of a single person (i.e. crop) within an image, not of the bag.

7.2.2 WL-MUDD Dataset

PerformancePhoto.co is an online marketplace for off-road racing photographers and fans. Powered by text spotting and ReID models to enable searchable racing photos, improvements to the ReID models suffer from the high costs of ReID dataset labeling. However, there is a proxy that gives natural weak labels: user purchases. When a user purchases photos from a single event, they are likely purchasing photos of a single individual. However, it is also likely that there is more than one individual in each photo purchased. Following notation from Section 7.2.1, the set of photos purchased by a single user can be regarded as a *bag* that can be weakly labeled with a unique identity.

The MUDD (Tyo et al., 2023) dataset was curated from PerformancePhoto.co and manually labeled in the traditional, fully supervised, ReID setting. We adapt the MUDD dataset to the weakly supervised setting by re-labeling the data points at the bag level and adding all, previously unlabeled, crops to each bag according to their existence in the original images. Figure 7.1 shows this labeling process and compares it to the standard (strong) annotation procedure. Instead of relying on the user purchases heuristic, we were able to build out the WL-MUDD dataset by taking all of the strong labels from the MUDD dataset, and then linking them back to the photos they originated from. Then, we take all the other people in the original photo, and add them to the dataset under the same label, forming a bag. This is repeated for every person in the original MUDD dataset, resulting in a weakly labeled dataset over twice as large. We refer to this dataset as the Weakly Labeled MUddy racer re-iDentification dataset (WL-MUDD).

The average bag in WL-MUDD has 75 crops of people in it, with 32% of them being the identity of the label attributed to that bag. This corresponds to an average noise level of 68%. The bags can be as small as 5 crops, or as large as 300, and the noise level of each bag varies between 50% and 85%. Figure 7.2 gives examples of bags in the dataset, highlighting the extremely high inter-class variation. The crops highlighted in green are representative of the bag label, whereas the red highlighted crops are not.

7.3 Contrastive Multiple Instance Learning

We cast the weakly supervised object re-identification problem as one of multiple-instance learning and present the contrastive multiple-instance learning (CMIL) method. A standard multiple-instance learning problem handles bag-level labeling by getting a feature representation for all crops in a bag, applying an accumulation function (typically max, average, etc.) to get

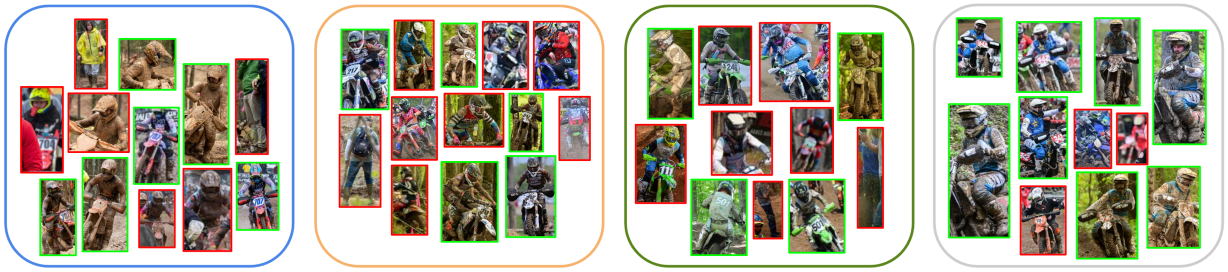


Figure 7.2: Four example subsets from four different bags of the WL-MUDD dataset. Each image within a bag is outlined in green if it is the same identity as the bag, and red if it is not. Each bag can have very different ratios of correct to incorrect identities of the underlying images.

a single bag representation from all of the crop representations, and then applying a classifier to the bag representation to determine a classification. Instead of a bag classifier (or alongside), we compare bag representations via a contrastive loss. This allows us to train end-to-end in a contrastive fashion. The CMIL framework is shown in Figure 7.3.

This is a divergence from standard contrastive learning. At test time we compare representations of crops, and therefore the goal of training is to optimize the crop representations accordingly. However, in this formulation, we are directly optimizing the bag representations, and only indirectly optimizing the crop representations. Specifically, all crops from a single bag k are encoded into crop representations by a model f parameterized by θ.

$$z_j^i = f_\theta(x_j^i),\ \forall j \in M_i, i \in N_k, \tag{7.1}$$

where M_i is the number of crops in image i and N_k is the set of images in bag k. It is critical that this model takes a specific crop as input, and returns the corresponding representation for that input because during testing, this is the only aspect of the model that will be utilized. Then given all crop representations for a bag k, they must be accumulated into a single bag-representation using a model g parameterized by ϕ.

$$r_k = g_\phi(z_1^1, \ldots, z_{M_i}^{N_k}). \tag{7.2}$$

This accumulation function should be permutation invariant to the input, as there is no way to control the ordering of the instances meaningfully.

The final component of this architecture is a distance or similarity function d. To apply contrastive learning, we must be able to measure the distances between pairs/triplets/quadruplets of bags. Any proper distance metric, such as the Euclidean or cosine distance, can be used. This distance can then be used to return a ranking, thresholded to provide a classification, etc. We focus on the setting where we are given a triplet of bags (or by the time it reaches the distance metric, bag representations). Given a bag a and b, the distance between their representations is represented by:

$$\hat{y} = d(r_a, r_b). \tag{7.3}$$

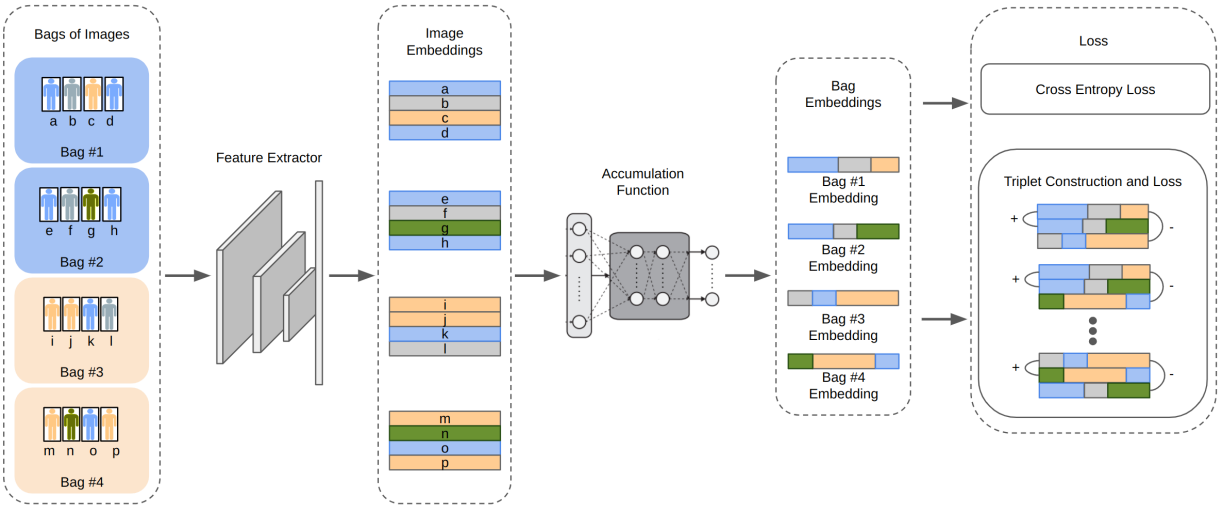


Figure 7.3: The CMIL framework. For each image in a batch of bags, a feature extraction network is used to get an embedding for each image. Then for each bag, the corresponding image embeddings are combined into a single bag embedding via an accumulation function. Finally, the bag embeddings are used to calculate the cross entropy loss (or identity loss), as well as the triplet loss based on all valid triplets from the batch.

Note that in this methodology, we depend on bags of data during each iteration. Each iteration must have a sufficient number of bags, as well as a sufficient number of crops from each bag. Therefore, the number of crops in each bag is intimately tied to both the batch size and the underlying assumptions about the nature (i.e. noise level) of the bags in the data. In most cases, the number of crops in a bag is large, and therefore we must sample mini-bags (e.g. a subset of a bag) to construct a batch. If the bag sizes are too small, then it is likely that there will not be enough of the true underlying identity in each bag to learn effectively. On the contrary, if the bags are too large, then it is likely that the number of bags in each batch is not sufficient for training. An implicit assumption of this framework is that in expectation, the most common identity in each bag is the identity representative of the bag label. The noise level can still be high without violating this assumption, because most non-representative crops in a bag are of different identities altogether. To ease notation, we will refer to bags and mini-bags interchangeably. In general, we mean mini-bags in algorithmic contexts, and bags in dataset contexts.

During inference, we follow the standard object re-identification procedure. Given a query set, gallery set, and an optional distractor set, we search for a specific object in the gallery based on a query image. All crops are embedded using our instance feature extractor, and then the distance metric used during training is used to return a ranking over the gallery and distractors for each query image. Based on this ranking, we track the rank-k accuracy for $k \in \{1, 5, 10\}$, as well as the mean average precision (mAP).

Algorithm 2 Contrastive Multiple Instance Learning (CMIL)

Require: Set of bags $\mathcal{B} = \{B_1, B_2, ..., B_N\}$, where each bag B_i contains crops $\{x_1^i, x_2^i, ..., x_{M_i}^i\}$
Ensure: Trained model parameters θ for crop feature extraction

Initialize model parameters θ, ϕ, and ψ randomly
while not converged **do**
  for $\mathcal{B}_{batch} \subset \mathcal{B}$ **do**
    for $B_i \in \mathcal{B}_{batch}$ **do**
      for $x_j^i \in B_i$ **do**
        $z_j^i = f_\theta(x_j^i)$ {Extract features from crops}
      end for
      $r_i = g_\phi(\{z_1^i, ..., z_{M_i}^i\})$ {Aggregate crop features into bag representation}
    end for
    $\mathcal{L}_{triplet}(r_{1,...,|B_{batch}|})$ {Triplet Loss}
    $\mathcal{L}_{CE}(h_\psi(r_{1,...,|B_{batch}|}))$ {CE Loss}
    $\mathcal{L}_{align}(r_{1,...,|B_{batch}|}, z^{1,...,|B_{batch}|}_{1,...,|M_i|})$ {Align Loss}
    $\mathcal{L} = \alpha\mathcal{L}_{triplet} + \beta\mathcal{L}_{CE} + \gamma\mathcal{L}_{align}$ {Aggregated Loss}
    Update model parameters θ, ϕ, ψ to minimize $\mathcal{L}$
  end for
end while

7.3.1 Loss Function

Table 7.1: Dataset Summary statistics for each dataset used in the experiments.

Dataset	Market-1501	SYSU-30k	weak MUDD
# identities	1,501	30,508	150
Scene	Outdoor	Indoor,Outdoor	Outdoor
Annotation	Strong	Weak	Weak
Cameras	6	Countless	Countless
Images	32,668	29,606,918	9,069

CMIL leverages both the identity and triplet losses. The identity loss is the cross entropy loss when each class represents a person identity

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{7.4}$$

where y_c is a binary indicator (0 or 1) indicating the label of a sample for class c, and p_c is the predicted probability of that class for the same sample, calculated by applying a fully connected layer (h parameterized by ψ) and a softmax to the bag representation:

$$p_i = \text{softmax}\Big(h_\psi(r_i)\Big). \tag{7.5}$$

Table 7.2: The sweep configuration for hyperparameter optimization, along with the final CMIL hyperparameters for each dataset. $U_{int}(x, y)$ represents an integer uniform distribution from x to y, $U_{log}(x, y)$ represents a log uniform, and $U(x, y)$ represents a standard uniform distribution on all real numbers from x to y.

Parameter	Search Range	Final Values		
		Market-1501	SYSU-30k	Weak MUDD
bag size	$U_{int}(5, 10)$	6	5	9
batch size	$U_{int}(5, 10)$	10	10	5
distance metric	[euclidean, cosine]	cosine	cosine	cosine
fixbase epoch	$U_{int}(0, 10)$	7	10	8
learning rate	$U_{log}(1e-05, 0.01)$	2.1153e-4	2.828e-3	4.044e-4
margin	$U(0.1, 1)$	0.9992	0.8592	0.7731
feature norm	[false, true]	False	False	False
gamma	[0, 0.01, 0.1]	0	0	0
alpha	$U(0, 1)$	0.5638	0.8083	0.3882
beta	$U(0, 1)$	0.3872	0.9242	0.7339

The triplet loss is:

$$\mathcal{L}_{\text{triplet}} = \max\Big(d(r_a, r_p) - d(r_a, r_n) + m_{\text{triplet}}, 0\Big) \tag{7.6}$$

where r_a is a bag representation for an anchor sample, r_p is a bag representation for a positive sample (i.e. a bag with the same label as the anchor sample), and r_n is a bag representation for a negative sample (i.e. a bag with a different label than the anchor sample).

Again, this is explicitly optimizing bag representations and only implicitly optimizing crop representations. In an attempt to address this, we experiment with an *alignment loss*. The intuition is that the most shared identity in a bag is the identity of interest. So an ideal accumulation function is one that can accurately pick out the representative crops, and then create a bag representation very similar to one or all of them. Therefore, we create the alignment loss to encourage the bag representation for a bag a with crops $\{x_1^a, x_2^a, \ldots, x_{N_a}^a\}$ to be close to any one of the crop representations:

$$\mathcal{L}_{\text{align}} = \max\Big(0, \min\{d(r_a, z_a^1), d(r_a, z_a^2), \ldots, d(r_a, z_a^{N_i})\} - m_{\text{align}}\Big), \tag{7.7}$$

where m_{align} is a margin hyperparameter.

The total loss function is a weighted combination of the identity, triplet, and alignment losses. The weighting for each loss (i.e. α, β, and γ) is selected during our hyperparameter search.

$$\mathcal{L} = \alpha\mathcal{L}_{\text{triplet}} + \beta\mathcal{L}_{\text{CE}} + \gamma\mathcal{L}_{\text{align}} \tag{7.8}$$

Note that the triplet loss can be substituted with any contrastive loss. Algorithm 2 provides an overview of CMIL in pseudocode.

Table 7.3: Results on the WL-Market1501 dataset at varying levels of noise. The noise level represents the percentage of the dataset with incorrect labels. This dataset was synthetically constructed by duplicating images in the training set and assigning them to random bags – 75% noise would correspond to duplicating each image three times, therefore only 1 in 4 images would be correctly labeled.

Noise	Method	R1	R5	R10	mAP
50%	CORE	**80.9%**	**92.2%**	**95.0%**	54.6%
	MIML	71.8%	87.0%	91.5%	46.3%
	CMIL (Ours)	80.7%	91.9%	94.4%	**56.8%**
66%	CORE	68.1%	83.6%	88.2%	38.6%
	MIML	62.7%	82.2%	87.6%	38.8%
	CMIL (Ours)	**76.4%**	**89.6%**	**93.0%**	**54.4%**
75%	CORE	56.1%	74.4%	80.8%	27.9%
	MIML	50.4%	71.6%	79.31%	26.6%
	CMIL (Ours)	**70.0%**	**86.4%**	**90.9%**	**48.8%**
80%	CORE	47.5%	66.0%	73.2%	17.4%
	MIML	54.0%	60.8%	71.1%	19.0%
	CMIL (Ours)	**64.9%**	**82.8%**	**88.0%**	**43.9%**

7.4 Experiments

We evaluate our methodology on three datasets:

- WL-Market-1501: The widely used Market-1501 person ReID dataset (Zheng et al., 2015), but with synthetically weak labels. The synthetic labels are generated by duplicating images from the training set some number of times, and assigning them to random bags.
- WL-MUDD: Our real-world dataset introduced in Section 7.2.2
- SYSU30k: A large-scale weakly supervised person ReID dataset with over 29 million images gathered from TV program videos. The videos are randomly broken into clips, and then each clip is manually annotated with an identity, but all detected people are noisily assigned that identity, forming bag-level labels.

The dataset statistics can be seen in Table 7.1. While the training set of each of these datasets is weakly labeled, the test sets are accurately labeled for normal person ReID evaluation (Ye et al., 2021b). We track the mean average precision (mAP) and the Rank-k accuracy for $k \in \{1, 5, 10\}$.

7.4.1 Implementation Details and Hyperparameter Tuning

7.4.2 Baseline Methods

Ye et al. (2021a) introduce online CO-REfining (CORE), a framework for online co-refining of ReID models. CORE uses learning rate schedules to optimize two models collaboratively, while also iteratively refining the noisy labels in a dataset. CORE is a state-of-the-art method for learning ReID models among noisy labels and weak supervision.

Meng et al. (2019) introduce Cross View Multi-Instance Multi-Label Learning (CV-MIML). Being based on MIL, this method falls most closely related to ours. Although originally developed for the setting where a target person is known to appear within an untrimmed video but no further information is available, this weakly supervised setting is equivalent to ours, although perhaps simpler due to correlations within a single video frame. Importantly, this method only performs bag classification during training, taking advantage only of the identity loss. Instead, CMIL optimizes bag representations explicitly.

We implement the CMIL framework using PyTorch. For a fair comparison, all methods utilize ResNet-50, pretrained on Imagenet (Deng et al., 2009), as the feature extractor f_θ. Our method also requires a reduction function g, and in this case, we use a 2-layer set transformer (Lee et al., 2019a). Section 7.5.1 includes ablations where we experiment with simpler reduction functions, namely the average, max, and sum operators. Importantly, we also implement a *bag* sampling function. We expect two conditions to be met for every mini-batch:

1. Each batch will consist of b sub-bags, where a sub-bag is a subset of a bag. If a bag is smaller than b, then the bag is oversampled.
2. Each bag label present in the mini-batch will have two or more bags in the mini-batch to ensure that valid triplets can always be constructed.

For hyperparameter selection, we run a Bayes hyperparameter search with early stopping (if model validation accuracy has not improved in 5 epochs, terminate the run) and hyperband (with an eta value of 2 and a minimum iteration count of 3) for early termination of less promising runs (Li et al., 2018a). Table 7.2 describes the hyperparameter search ranges. The search aims to maximize the rank-1 accuracy on the validation set over 50 epochs. For each dataset, 250 models with hyperparameters sampled from the listed distributions were trained and evaluated, and the best-performing hyperparameters are also shown in Table 7.2. Finally, using the best-performing hyperparameters, a final training run was done using the combined training and validation set, evaluated on the test set, and reported in our results.

7.5 Results and Discussion

Table 7.3 summarizes the rank-1 (R1), rank-5 (R5), rank-10 (R10) accuracy and mean average precision (mAP) of the different methods on the Market-1501 dataset with varying levels of synthetic label noise. At the 50% noise level, our CMIL method achieves an R1 accuracy of 80.7%, nearly matching the performance of CORE and outperforming MIML by 8.9%. As the noise level begins to increase, CMIL dominates the other methods by a growing margin. At 80% label noise, the hardest setting, CMIL obtains a R1 of 64.9%, which signifies a 10.9% boost over the best baseline. The consistent gaps between CMIL and other approaches illustrate that our method can effectively learn useful representations among even extremely noisy bags.

Table 7.4 summarizes the performance of different methods when trained on the real-world Weak MUDD dataset. With noisy group annotations, our CMIL framework obtains 73.2% rank-1 accuracy. This significantly outperforms baseline methods, including CORE and MIML, by 2.5% and 6% respectively.

Table 7.4: Results on the WL-MUDD dataset.

Method	R1	R5	R10	mAP
CORE	67.2%	83.3%	92.3%	71.6%
MIML	70.7%	87.7%	95.2%	74.6%
CMIL (Ours)	**73.2%**	**90.0%**	**96.8%**	**75.1%**

Table 7.5: Results on the SYSU30k dataset.

Supervision	Method	R1
Transfer Learning	DARIR (Wang et al., 2016a)	11.2%
	DF (Ding et al., 2015)	10.3%
	Local CNN (Yang et al., 2018a)	23.0%
	MGN (Wang et al., 2018b)	23.6%
Self-Supervised	SimCLR (Chen et al., 2020a)	10.9%
	MoCo v2 (Chen et al., 2020b)	11.6%
	BYOL (Grill et al., 2020)	12.7%
	Triplet (Wang et al., 2021a)	27.5%
Weakly Supervised	W-Local CNN (Wang et al., 2020a)	28.8%
	W-MGN (Wang et al., 2020a)	29.5%
	WS-TAL (Liu et al., 2023)	**34.4%**
	CMIL (Ours)	33.9%

The SYSU-30k dataset is very large and computationally expensive to optimize models on. Therefore, we compare directly to the results reported in prior work in Table 7.5. CMIL attains

33.9% R1 accuracy outperforming the best transfer and self-supervised learning approaches by 10.3% and 6.4% respectively. The best weakly supervised method is WS-TAL (Liu et al., 2023), which is specifically engineered to optimize ReID models when the labels are generated from video tracklets, matching the SYSU construction. WS-TAL reaches 34.4% R1 accuracy. CMIL nearly matches this performance, lagging by only 0.5%, using more general labeling assumptions.

Surprisingly, in every case, the alignment loss does not improve accuracy – as shown in Table 7.2, $\gamma = 0$ and therefore the alignment loss is not used. During training, CMIL models are optimized at the bag level. Given a batch of bag representations, the model is optimized for bags with the same label to be close, and bags with a different label to be far from each other in representation space. The bag representations are built from crop representations, but nothing is preventing the bag and crop representations from being far apart. This should be problematic, because at test time, we are evaluating the quality of the crop embeddings.

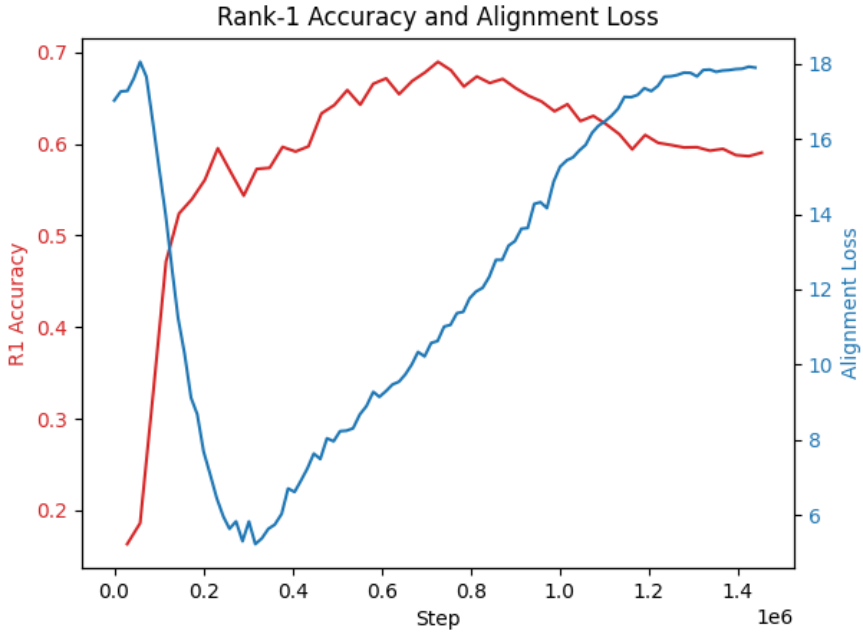


Figure 7.4: The rank-1 accuracy and the alignment loss throughout a training run. The alignment loss exhibits unintuitive behavior - the best alignment (i.e. lowest) does not correspond to the best model accuracy (i.e. highest). This behavior is characteristic of every model trained in this work, including those using different accumulation functions.

Figure 7.4 plots the rank-1 accuracy and alignment loss versus training step when $\gamma = 0$ (i.e. the alignment loss is not used). We see that the bag and crop representations start quite different, and then begin growing more similar. However, there reaches a point relatively early in training where the alignment between instance and bag representations begins to decrease. Interestingly, the performance of the model (therefore the *crop* embedding model) continues to improve, even while their embeddings diverge from bag embeddings. This phenomenon is

Table 7.6: Comparison of different accumulation functions on WL-Market-1501 and WL-MUDD datasets. Using a simple average of crop representations performs nearly as well as the set transformer.

	WL-Market1501				WL-MUDD			
Method	R1	R5	R10	mAP	R1	R5	R10	mAP
CMIL w/Set Transformer	70.0%	86.4%	90.9%	48.8%	73.2%	90.0%	96.8%	75.1%
CMIL w/Max	60.2%	78.7%	84.6%	33.9%	66.8%	82.1%	90.7%	68.2%
CMIL w/Avg	69.8%	85.6%	90.1%	44.1%	71.1%	88.6%	94.8%	74.5%
CMIL w/Sum	51.2%	72.0%	79.6%	24.3%	64.3%	79.3%	88.9%	62.4%

consistent across all of our experiments.

It is not likely that this counterintuitive behavior is an artifact of the way we are measuring the alignment loss. It is not known which crop within a bag is the crop representative of the bag label, so the alignment loss used here is the distance, using the same measure used during training, between the bag representation and the *closest* crop representation within the bag. A better understanding of this phenomenon requires further research.

7.5.1 Ablation Study

In this ablation, we experiment with other, more traditional, choices for permutation invariant accumulation function, specifically the max, average, and sum. Each bag contains crops, and one or more of the crops in the bag are representative of the bag label. Of course, which crop specifically is unknown. Intuitively, the job of the accumulation function is to select the crop (or a representation of the collection of crops) that corresponds to the bag label.

All aforementioned models have used a set transformer for the accumulation function, as the learnable attention-based model makes it possible to behave as a selector, or any arbitrary combination of the crop representations. However, it does come at the cost of complexity. Other reasonable, and much simpler, choices are to set the bag representation to be the max, average, or sum of the crop representations.

In Table 7.6, we compare the performance of the different accumulation functions on both the WL-Market1501 and the WL-MUDD datasets. Interestingly, the average does well, matching, or nearly matching, the performance of the set transformer. This is surprising as the crops within a representative bag of the bag label are the minority of samples, typically representing less than half of the samples within each bag. This could indicate that the set transformer is roughly just performing an average. A potential reason for this is that the non-representative crop features could act to cancel one another out such that the bag representation is still close to the corresponding representative features. Interestingly, tracking the alignment loss for each of these simpler accumulation functions shows the exact behavior as depicted in Figure 7.4.

7.6 Related Work

A large body of work has focused on supervised re-ID, where models are trained on data with individual object identity labels (Ye et al., 2021b; Zheng et al., 2016; Zheng et al., 2017; Li et al., 2018c; He et al., 2021). These approaches employ deep neural networks, to extract visual features that are representative of specific identities, and optimize them according to various loss functions, including the identification loss where each identity is treated as a class (Zheng et al., 2017), verification loss where pairwise relationships are optimized vi the contrastive loss (Varior et al., 2016), triplet loss that treats the problem as a retrieval ranking problem (Hermans et al., 2017), and others have been proposed to optimize re-ID performance (Chen et al., 2017; Yang et al., 2019; Liu et al., 2019; Song et al., 2019; Wang et al., 2018a; Chen et al., 2019a; Zhou et al., 2019b; Xiao et al., 2017; Guo et al., 2019). These methods perform well on baseline datasets, where ample data labeling is available.

To alleviate the labeling bottleneck, recent works have begun investigating re-ID under weak supervision. Strategies include exploiting image-level labels (Meng et al., 2019), pseudo-labels (Wang et al., 2020a), noisy label refinement (Ye et al., 2021a), online captions (Zhao et al., 2021; Guillaumin et al., 2010), and domain adaptation (Yu et al., 2023). While showing promise, these methods still fall behind those that are fully supervised (Zheng et al., 2021).

Most similarly to our work, Meng et al. (2019) leverage image-level labels in conjunction with Multiple Instance Learning (MIL) to effective facial recognition models. MIL offers a paradigm to handle label ambiguity in training data by modeling labels at a bag level. A bag can be a collection of instances associated with a particular label, but we only know that one or more of the instances in that collection truly belongs to that label. Several works have adapted this specifically for treating video "tracklets" as a bag of instances (Liu et al., 2023; Wang et al., 2021b) MIL has found diverse applications including image classification (particularly medical imagery) (Wu et al., 2015; Sudharshan et al., 2019), object detection (Yuan et al., 2021; Wan et al., 2019; Huang et al., 2017), and drug discovery (Fu et al., 2012). Critically, Meng et al. (2019) apply MIL to the person re-identification problem in the identity loss setting. In contrast, CMIL improves upon this by allowing for use of contrastive learning, which has shown significant advantages in person ReID and related settings (Hermans et al., 2017; Garg et al., 2023).

Lastly, we must mention the work in unsupervised ReID (Fu et al., 2021; Li et al., 2018b; Wang and Zhang, 2020; Lin et al., 2019; Yu et al., 2019; Fan et al., 2018). These methods do not require labels. Typically, these methods use iterative clustering and classification, such that unlabeled images are clustered into "pseudo" classes, which are then used to train or update a model. Then the new/updated model is used to refine the pseudo labels, and so on. Improvements to this standard approach include substituting the clustering step for pairwise comparisions (Lin et al., 2020), and an improved clustering step by improving the global clusters using ensembles of image-part based predictions (Cho et al., 2022). Of course, performance is still greatly improved when labels are present (Xiang et al., 2023; Zhu et al., 2022; Luo et al., 2021; Yang et al., 2022; Fu et al., 2022).

7.7 Conclusion and Future Work

In this paper, we introduced Contrastive Multiple Instance Learning (CMIL), a novel framework tailored for more effective person re-identification under weak supervision. CMIL tackles the challenge of learning discriminative person representations when only bag-level labels indicating a shared identity among a group of photos are available. Although the model is trained at the bag level, the person-level representations improve alongside the quality of the bag-level representations. We experiment with adding an alignment loss to further encourage the person and bag representations to be similar, but found it ineffective empirically.

We experiment on three datasets, one of which is the Weakly Labeled Muddy Racer Re-Identification Dataset (WL-MUDD), which is curated and released from real-world weak labels from PerformancePhoto.co. Across these experiments, CMIL consistently achieved state-of-the-art rank-1, rank-5, and rank-10 accuracy as well as mean average precision. On the large-scale SYSU-30k dataset, CMIL matched the top-reported result while requiring fewer assumptions. Ablations also revealed surprising effectiveness of average pooling for instance aggregation, suffering only slight performance degradation to the set transformer.

The contributions of this work are threefold. First, we introduce the new Weakly Labeled Muddy Racer Re-Identification dataset (WL-MUDD) built from PerformancePhoto.co, an off-road photograph platform. Second, we introduce the CMIL framework that enables efficient exploitation of cheap weak supervision for person re-id through enabling contrastive learning with Multiple Instance Learning. And third, we show the efficacy of CMIL on two real-world datasets and one synthetic, outperforming baselines.

We note that in the creation of bags, the information about crop comes from what image is lost. This information can potentially improve the performance of such a system notably, especially in the case where there are some images that contain only a single person, and therefore we would know that is the person of interest. A good direction for future work is determining how to incorporate this information correctly, giving the model more meaninful grouding during training.

Part III

Conclusion

Chapter 8

Conclusion

This thesis studies the practical application of contrastive learning and develops new methods for improving performance in these real-world settings. In particular, it is driven by two questions:

1. *How can we understand the practical benefits of contrastive versus non-contrastive models?*
2. *How can we refine contrastive methods to integrate the strengths of both paradigms?*

Part I is dedicated to making progress on the first question. Through examination of contrastive learning in practical scenarios, including authorship attribution, authorship verification, and person re-identification, the adaptability of contrastive learning techniques proved a strength that can be augmented with an appropriate mining function for maximal contrastive performance, and also compete with non-contrastive methods in non-contrastive settings. Part II is dedicated to leveraging those findings to improve the underlying learning procedures themselves. Through the introduction of meta-learning to enable learning intelligent sample selection functions (i.e. the mining function), and through the bridging of contrastive learning with multiple instance learning, much stronger contrastive models can be trained.

Chapter 2 highlights how contrastive models excel over classification models any time a distribution shift is induced on the testing set. Chapter 3.3 backs up these results in many other settings and further indicates that assuming the use of hard-negative mining, contrastive models can perform equally to a classification model in a classification setting, while still maintaining the advantages of being more robust to domain shift, flexible for contrastive and retrieval settings, and able to perform some level of zero or few-shot learning without any gradient updates. Furthermore, Chapter 3.3 introduces the VALLA benchmark, complete with method implementations, data loaders, and full training pipelines, for accurately benchmarking authorship identification methods, including the classification vs the contrastive methodologies, apples-to-apples. Focusing further on the importance of hard negative mining as indicated in Chatper 3.3, Chapter 6 turns away from the hand-engineered nature of hard-negative mining

and towards meta-learning to enable the automatic discovery of such mining functions. The learned mining functions match the performance of the best hand-engineered function, but it removes the uncertainty of trying to determine the right hand-engineered mining function and can discover new mining functions if necessary.

Chapters 4 and 5 introduce, release, and benchmark models on new datasets in the off-road racing domain. While efforts at developing a new data augmentation strategy did lead to improved performance, the major effort for improving performance was derived from the observation that an abundance of data was available, but labeled only extremely weakly (¿50% label noise). To address this, Chapter 7 introduced a multiple-instance learning-based methodology that enabled effective model training even among this extreme amount of noise.

Real-world deployments are typically fraught with unpredictability and data variability, as emphasized by the off-road datasets. The methods introduced in this thesis take steps in building powerful models that excel, even in extremely unfriendly scenarios. Along with highlighting the adaptability of contrastive learning techniques in overcoming real-world challenges, this work shows that contrastive models excel over classification models any time a distribution shift is induced on the testing set. Furthermore, assuming the use of hard-negative mining (or learning a mining function), contrastive models can perform equally to a classification model in a classification setting, while maintaining the advantages of being more robust to domain shift, flexible for contrastive and retrieval settings, and able to perform some level of zero or few-shot learning without any gradient updates. Finally, even in extremely noisy settings where normal contrastive methods fail, we can still realize the benefits of contrastive models via contrastive multiple-instance learning.

This thesis represents a comprehensive effort to advance the understanding and application of contrastive learning in machine learning. While this journey has illuminated numerous facets of contrastive learning, it has also underscored the vast expanse of uncharted territory that remains. I hope that this work will catalyze future research, inspiring continued exploration and innovation in the ever-evolving landscape of machine learning, especially among difficult niche applications.

8.1 Future Work

This thesis has laid a substantial foundation in the field of contrastive learning, addressing various applications and methodological challenges. However, as with any exploratory work, the completion of this research opens new and promising avenues for further investigation.

There remains significant untapped potential in the realm of authorship verification, particularly regarding the utilization of the latest generation of large pretrained models. Future research could leverage the Valla environment to benchmark these advanced NLP models comprehensively. Additionally, the development of methodologies that are more robust to noise

could enhance the accuracy and reliability of authorship verification, especially in an era where the definition of authorship is becoming increasingly blurred by the capabilities of LLMs.

Despite our advancements in data augmentation techniques for OCR of off-road racer numbers, there is still a considerable gap between current model performance and the accuracy levels achieved in structured document understanding or by human operators. Addressing this gap represents a significant opportunity for further research, potentially through the exploration of novel data augmentation strategies better tailored for the color changes induced by mud, glare, shadows, or dust or the development of more sophisticated models tailored to the unique challenges of this application. Accuracy remains low yielding this as a fruitful research area.

While progress has been made in person re-identification for off-road racing contexts, challenges remain, particularly related to the distribution shift caused by racers becoming covered in mud. Future work could explore additional data augmentation techniques that more accurately model the changes in appearance due to mud and other environmental factors. Investigating new strategies for data sampling, such as pairing images of the same individual under drastically different conditions, may also improve model robustness.

The development of risk-adjusted mini-batches has shown promise in providing a better approximation of complex risk measures at the mini-batch level. However, these approaches still suffer from bias and are slow and resource-intensive to optimize. Future research should aim to develop methods that can either eliminate this bias or optimize models more quickly and with reduced memory requirements, thereby enhancing the practical utility of these approaches. Furthermore, because of the long computational graphs, the gradients can be quite unstable, rendering medium to high learning rates unusable. Being forced into using only very small learning rates is also a hindrance that can be improved.

The current methodologies in contrastive multiple-instance learning focus on optimizing bag representations, with the assumption that improving the bag will inherently enhance individual representations. However, this assumption may not always hold true. Therefore, a critical area for future investigation is the development of strategies that explicitly optimize individual representations within the framework of multiple-instance learning. This could involve the creation of novel loss functions or optimization techniques that ensure improvements at the individual level are explicitly targeted and achieved.

Bibliography

[1] Lorenzo Valla. "Discourse on the Forgery of the Alleged Donation of Constantine". In: *Latin and English. Trans. Christopher B. Coleman. New Haven* (1922) (cit. on p. 28).

[2] Frederick Mosteller and David L. Wallace. "Inference in an Authorship Problem". In: *Journal of the American Statistical Association* 58.302 (1963), pp. 275–309. DOI: 10.1080/01621459.1963.10500849. eprint: https://doi.org/10.1080/01621459.1963.10500849. URL: https://doi.org/10.1080/01621459.1963.10500849 (cit. on p. 21).

[3] Dmitri V Khmelev and Fiona J Tweedie. "Using Markov chains for identification of writer". In: *Literary and linguistic computing* 16.3 (2001), pp. 299–307 (cit. on p. 25).

[4] William J. Teahan and David J. Harper. "Using Compression-Based Language Models for Text Categorization". In: *Language Modeling for Information Retrieval*. Ed. by W. Bruce Croft and John Lafferty. Dordrecht: Springer Netherlands, 2003, pp. 141–165. ISBN: 978-94-017-0171-6. DOI: 10.1007/978-94-017-0171-6_7. URL: https://doi.org/10.1007/978-94-017-0171-6_7 (cit. on p. 29).

[5] Bryan Klimt and Yiming Yang. "The enron corpus: A new dataset for email classification research". In: *European conference on machine learning*. Springer. 2004, pp. 217–226 (cit. on p. 28).

[6] Moshe Koppel and Jonathan Schler. "Authorship verification as a one-class classification problem". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 62 (cit. on p. 26).

[7] Hsi-Jian Lee, Si-Yuan Chen, and Shen-Zheng Wang. "Extraction and recognition of license plates of motorcycles and vehicles on highways". In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 4. IEEE. 2004, pp. 356–359 (cit. on p. 38).

[8] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. "Rcv1: A new benchmark collection for text categorization research". In: *Journal of machine learning research* 5.Apr (2004), pp. 361–397 (cit. on p. 24).

[9] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. "Effects of age and gender on blogging." In: *AAAI spring symposium: Computational approaches to analyzing weblogs*. Vol. 6. 2006, pp. 199–205 (cit. on p. 24).

[10] Jade Goldstein, Kerri Goodwin, Roberta Sabin, and Ransom Winder. "Creating and Using a Correlated Corpus to Glean Communicative Commonalities". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. 2008 (cit. on p. 24).

[11] Moshe Koppel, Jonathan Schler, and Eran Messeri. "Authorship attribution in law enforcement scenarios". In: *NATO Security Through Science Series D-Information and Communication Security* 15 (2008), p. 111 (cit. on p. 21).

[12] Matko Šaric, Hrvoje Dujmic, Vladan Papic, and Nikola Rožic. "Player number localization and recognition in soccer video using hsv color space and internal contours". In: *International Journal of Electrical and Computer Engineering* 2.7 (2008), pp. 1408–1412 (cit. on p. 39).

[13] Rajkumar Arun, Venkatasubramaniyan Suresh, and CE Veni Madhavan. "Stopword graphs and authorship attribution in text corpora". In: *2009 IEEE international conference on semantic computing*. IEEE. 2009, pp. 192–196 (cit. on pp. 25, 26).

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on pp. 52, 97).

[15] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009) (cit. on pp. 72, 76).

[16] William Robson Schwartz and Larry S Davis. "Learning discriminative appearance-based models using partial least squares". In: *2009 XXII Brazilian symposium on computer graphics and image processing*. IEEE. 2009, pp. 322–329 (cit. on p. 60).

[17] Efstathios Stamatatos. "A survey of modern authorship attribution methods". In: *Journal of the American Society for Information Science and Technology* 60.3 (2009), pp. 538–556. DOI: https://doi.org/10.1002/asi.21001. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21001. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001 (cit. on p. 23).

[18] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. "Person re-identification by symmetry-driven accumulation of local features". In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 2360–2367 (cit. on p. 60).

[19] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. "Multiple instance metric learning from automatically labeled bags of faces". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*. Springer. 2010, pp. 634–647 (cit. on pp. 88, 101).

[20] Stuart J Russell and Peter Norvig. *Artificial intelligence a modern approach*. London, 2010 (cit. on p. 1).

[21] Kai Wang and Serge Belongie. "Word spotting in the wild". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*. Springer. 2010, pp. 591–604 (cit. on p. 38).

[22] Ana Granados, Manuel Cebrian, David Camacho, and Francisco de Borja Rodriguez. "Reducing the Loss of Information through Annealing Text Distortion". In: *IEEE Trans-*

actions on Knowledge and Data Engineering 23.7 (2011), pp. 1090–1102. DOI: 10.1109/TKDE.2010.173 (cit. on p. 26).

[23] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. "Authorship attribution in the wild". In: *Language Resources and Evaluation* 45.1 (2011), pp. 83–94 (cit. on pp. 24, 26).

[24] Rohith Menon and Yejin Choi. "Domain independent authorship attribution without domain adaptation". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. 2011, pp. 309–315 (cit. on p. 25).

[25] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. "Reading digits in natural images with unsupervised feature learning". In: (2011) (cit. on p. 36).

[26] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on p. 15).

[27] Anselmo Peñas and Alvaro Rodrigo. "A simple measure to assess non-response". In: (2011) (cit. on p. 15).

[28] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. "Authorship attribution with latent Dirichlet allocation". In: *Proceedings of the fifteenth conference on computational natural language learning*. 2011, pp. 181–189 (cit. on p. 26).

[29] Efstathios Stamatatos and Moshe Koppel. "Plagiarism and authorship analysis: introduction to the special issue". In: *Language Resources and Evaluation* 45.1 (2011), pp. 1–4 (cit. on p. 21).

[30] Kai Wang, Boris Babenko, and Serge Belongie. "End-to-end scene text recognition". In: *2011 International conference on computer vision*. IEEE. 2011, pp. 1457–1464 (cit. on p. 38).

[31] Simi Wang, Michal Lewandowski, James Annesley, and James Orwell. "Re-identification of pedestrians with variable occlusion and scale". In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE. 2011, pp. 1876–1882 (cit. on p. 60).

[32] Idan Ben-Ami, Tali Basha, and Shai Avidan. "Racing Bib Numbers Recognition." In: *BMVC*. 2012, pp. 1–10 (cit. on p. 39).

[33] Gang Fu, Xiaofei Nan, Haining Liu, Ronak Y Patel, Pankaj R Daga, Yixin Chen, Dawn E Wilkins, and Robert J Doerksen. "Implementation of multiple-instance learning in drug activity prediction". In: *BMC bioinformatics*. Vol. 13. 15. BioMed Central. 2012, pp. 1–12 (cit. on p. 101).

[34] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. "Detecting texts of arbitrary orientations in natural images". In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 1083–1090 (cit. on p. 44).

[35] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. "ICDAR 2013 robust reading competition". In: *2013 12th international conference on document analysis and recognition*. IEEE. 2013, pp. 1484–1493 (cit. on pp. 38, 43, 44).

[36] Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. "Consumer credit risk: Individual probability estimates using machine learning". In: *Expert systems with applications* 40.13 (2013), pp. 5125–5131 (cit. on p. 71).

[37] Stefano Messelodi and Carla Maria Modena. "Scene text recognition and tracking to identify athletes in sport videos". In: *Multimedia tools and applications* 63.2 (2013), pp. 521–545 (cit. on p. 39).

[38] Jacques Savoy. "Authorship attribution based on a probabilistic topic model". In: *Information Processing & Management* 49.1 (2013), pp. 341–354 (cit. on p. 11).

[39] Efstathios Stamatatos. "On the robustness of authorship attribution based on character n-gram features". In: *Journal of Law and Policy* 21.2 (2013), pp. 421–439 (cit. on p. 24).

[40] Taiki Yamamoto, Hirokatsu Kataoka, Masaki Hayashi, Yoshimitsu Aoki, Kyoko Oshima, and Masamoto Tanabiki. "Multiple players tracking and identification using group detection and player number recognition in sports video". In: *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*. IEEE. 2013, pp. 2442–2446 (cit. on p. 39).

[41] Tomas Björk, Agatha Murgoci, and Xun Yu Zhou. "Mean–variance portfolio optimization with state-dependent risk aversion". In: *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* 24.1 (2014), pp. 1–24 (cit. on p. 73).

[42] Sara El Manar El Bouanani and Ismail Kassou. "Authorship analysis studies: A survey". In: *International Journal of Computer Applications* 86.12 (2014) (cit. on p. 23).

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755 (cit. on p. 38).

[44] Naruemon Pratanwanich and Pietro Lio. "Who wrote this? Textual modeling with authorship attribution in big data". In: *2014 IEEE International Conference on Data Mining Workshop*. IEEE. 2014, pp. 645–652 (cit. on p. 25).

[45] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. "Authorship attribution with topic models". In: *Computational Linguistics* 40.2 (2014), pp. 269–310 (cit. on pp. 24, 26).

[46] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014 (cit. on p. 1).

[47] Min Yang and Kam-Pui Chow. "Authorship attribution for forensic investigation with thousands of authors". In: *IFIP International Information Security Conference*. Springer. 2014, pp. 339–350 (cit. on p. 24).

[48] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. "Deep metric learning for person re-identification". In: *2014 22nd international conference on pattern recognition*. IEEE. 2014, pp. 34–39 (cit. on p. 61).

[49] Douglas Bagnall. "Author identification using multi-headed recurrent neural networks". In: *arXiv preprint arXiv:1506.04891* (2015) (cit. on pp. 11, 25, 27).

[50] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. "Deep feature learning with relative distance comparison for person re-identification". In: *Pattern Recognition* 48.10 (2015), pp. 2993–3003 (cit. on p. 98).

[51] Dario Figueira, Matteo Taiana, Athira Nambiar, Jacinto Nascimento, and Alexandre Bernardino. "The HDA+ data set for research on fully automated re-identification systems". In: *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part III 13*. Springer. 2015, pp. 241–255 (cit. on p. 60).

[52] Sebastian Gerke, Karsten Muller, and Ralf Schafer. "Soccer Jersey Number Recognition Using Convolutional Neural Networks". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*. 2015 (cit. on p. 39).

[53] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. "ICDAR 2015 competition on robust reading". In: *2015 13th international conference on document analysis and recognition (ICDAR)*. IEEE. 2015, pp. 1156–1160 (cit. on p. 43).

[54] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. "Siamese neural networks for one-shot image recognition". In: *ICML deep learning workshop*. Vol. 2. Lille. 2015, p. 0 (cit. on p. 27).

[55] Tongliang Liu and Dacheng Tao. "Classification with noisy labels by importance reweighting". In: *IEEE Transactions on pattern analysis and machine intelligence* 38.3 (2015), pp. 447–461 (cit. on p. 73).

[56] Dylan Rhodes. "Author attribution with cnns". In: *Avaiable online: https://www.semanticscholar.org/paper Attribution-with-Cnn-s-Rhodes/ (accessed on 22 August 2016)* (2015) (cit. on p. 25).

[57] Zhiyuan Shi, Timothy M Hospedales, and Tao Xiang. "Transferring a semantic representation for person re-identification and search". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4184–4193 (cit. on p. 61).

[58] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. "Deep multiple instance learning for image classification and auto-annotation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3460–3469 (cit. on pp. 88, 101).

[59] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. "Learning from massive noisy labeled data for image classification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2691–2699 (cit. on p. 73).

[60] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. "Scalable person re-identification: A benchmark". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1116–1124 (cit. on pp. 58, 60, 83, 89, 96).

[61] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. "Person re-identification by multi-channel parts-based cnn with improved triplet loss function". In: *Proceedings of the iEEE conference on computer vision and pattern recognition*. 2016, pp. 1335–1344 (cit. on p. 61).

[62] Yeong-Jun Cho and Kuk-Jin Yoon. "Improving person re-identification via pose-aware multi-shot matching". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1354–1362 (cit. on p. 61).

[63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 43, 52, 56, 57, 76, 89).

[64] Rebekah Overdorf and Rachel Greenstadt. “Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution.” In: *Proc. Priv. Enhancing Technol.* 2016.3 (2016), pp. 155–171 (cit. on p. 28).

[65] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. “Performance measures and a data set for multi-target, multi-camera tracking”. In: *European conference on computer vision*. Springer. 2016, pp. 17–35 (cit. on pp. 58, 60).

[66] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. “Character-level and multi-channel convolutional neural networks for large-scale authorship attribution”. In: *arXiv preprint arXiv:1609.06686* (2016) (cit. on pp. 11, 24, 27).

[67] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. “Detecting text in natural image with connectionist text proposal network”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer. 2016, pp. 56–72 (cit. on p. 38).

[68] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. “A siamese long short-term memory architecture for human re-identification”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer. 2016, pp. 135–153 (cit. on p. 101).

[69] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. “Coco-text: Dataset and benchmark for text detection and recognition in natural images”. In: *arXiv preprint arXiv:1601.07140* (2016) (cit. on p. 44).

[70] Guangrun Wang, Liang Lin, Shengyong Ding, Ya Li, and Qing Wang. “DARI: Distance Metric and Representation Integration for Person Verification”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 30. 1. 2016 (cit. on p. 98).

[71] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. “Person re-identification by discriminative selection in video ranking”. In: *IEEE transactions on pattern analysis and machine intelligence* 38.12 (2016), pp. 2501–2514 (cit. on p. 60).

[72] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. “Google’s neural machine translation system: Bridging the gap between human and machine translation”. In: *arXiv preprint arXiv:1609.08144* (2016) (cit. on p. 12).

[73] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. “End-to-end deep learning for person search”. In: *arXiv preprint arXiv:1604.01850* 2.2 (2016), p. 4 (cit. on p. 60).

[74] Liang Zheng, Yi Yang, and Alexander G Hauptmann. “Person re-identification: Past, present and future”. In: *arXiv preprint arXiv:1610.02984* (2016) (cit. on pp. 60, 101).

[75] Sylvio Barbon, Rodrigo Augusto Igawa, and Bruno Bogaz Zarpelão. “Authorship verification applied to detection of compromised accounts on online social networks”. In: *Multimedia Tools and Applications* 76.3 (2017), pp. 3213–3233 (cit. on p. 21).

[76] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. “Active bias: Training more accurate neural networks by emphasizing high variance samples”. In: *Advances in Neural Information Processing Systems* 30 (2017) (cit. on p. 73).

[77] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. "Beyond triplet loss: a deep quadruplet network for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 403–412 (cit. on p. 101).

[78] Chee Kheng Ch'ng and Chee Seng Chan. "Total-text: A comprehensive dataset for scene text detection and recognition". In: *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 935–942 (cit. on p. 44).

[79] Sebastian Gerke, Antje Linnemann, and Karsten Müller. "Soccer player recognition using spatial constellation features and jersey number recognition". In: *Computer Vision and Image Understanding* 159 (2017), pp. 105–115 (cit. on p. 39).

[80] Raul Gomez, Baoguang Shi, Lluis Gomez, Lukas Numann, Andreas Veit, Jiri Matas, Serge Belongie, and Dimosthenis Karatzas. "Icdar2017 robust reading challenge on coco-text". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 1435–1443 (cit. on pp. 43, 44).

[81] Oren Halvani, Christian Winter, and Lukas Graner. "On the usefulness of compression models for authorship verification". In: *Proceedings of the 12th international conference on availability, reliability and security*. 2017, pp. 1–10 (cit. on pp. 24, 27).

[82] Alexander Hermans, Lucas Beyer, and Bastian Leibe. "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737* (2017) (cit. on pp. 12, 15, 33, 88, 101).

[83] Fang Huang, Jinqing Qi, Huchuan Lu, Lihe Zhang, and Xiang Ruan. "Salient object detection via multiple instance learning". In: *IEEE Transactions on Image Processing* 26.4 (2017), pp. 1911–1922 (cit. on pp. 88, 101).

[84] Kamlesh, Pei Xu, Yang Yang, and Yongchao Xu. "Person re-identification with end-to-end scene text recognition". In: *Computer Vision: Second CCF Chinese Conference, CCCV 2017, Tianjin, China, October 11–14, 2017, Proceedings, Part III*. Springer. 2017, pp. 363–374 (cit. on p. 39).

[85] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. "Surveying Stylometry Techniques and Applications". In: *ACM Comput. Surv.* 50.6 (2017). ISSN: 0360-0300. DOI: 10.1145/3132039. URL: https://doi.org/10.1145/3132039 (cit. on pp. 22, 23, 29).

[86] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. "Making deep neural networks robust to label noise: A loss correction approach". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1944–1952 (cit. on p. 73).

[87] Palaiahnakote Shivakumara, Ramachandra Raghavendra, Longfei Qin, Kiran B Raja, Tong Lu, and Umapada Pal. "A new multi-modal approach to bib number/text detection and recognition in Marathon images". In: *pattern recognition* 61 (2017), pp. 479–491 (cit. on p. 38).

[88] Prasha Shrestha, Sebastian Sierra, Fabio A González, Manuel Montes, Paolo Rosso, and Thamar Solorio. "Convolutional neural networks for authorship attribution of short texts". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 2017, pp. 669–674 (cit. on pp. 11, 27).

[89] Ruxin Wang, Tongliang Liu, and Dacheng Tao. "Multiclass learning with partially corrupted labels". In: *IEEE transactions on neural networks and learning systems* 29.6 (2017), pp. 2568–2580 (cit. on p. 73).

[90] Ada Wrońska, Kacper Sarnacki, and Khalid Saeed. "Athlete number detection on the basis of their face images". In: *2017 International Conference on Biometrics and Kansei Engineering (ICBAKE)*. IEEE. 2017, pp. 84–89 (cit. on p. 39).

[91] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. "Joint detection and identification feature learning for person search". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3415–3424 (cit. on p. 101).

[92] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. "Person re-identification in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1367–1376 (cit. on p. 101).

[93] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. "EAST: An Efficient and Accurate Scene Text Detector". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on p. 38).

[94] Zubaer Ahammed. "Basketball player identification by jersey and number recognition". PhD thesis. Brac University, 2018 (cit. on p. 39).

[95] Bushra Alhijawi, Safaa Hriez, and Arafat Awajan. "Text-based authorship identification-A survey". In: *2018 Fifth International Symposium on Innovation in Information and Communication Technology (ISIICT)*. IEEE. 2018, pp. 1–7 (cit. on p. 24).

[96] Shlomo Argamon. "Computational forensic authorship analysis: Promises and pitfalls". In: *Language and Law/Linguagem e Direito* 5.2 (2018), pp. 7–37 (cit. on p. 24).

[97] Noppakun Boonsim. "Racing bib number localization on complex backgrounds". In: *WSEAS Transactions on Systems and Control* 13 (2018), pp. 226–231 (cit. on p. 39).

[98] Carolina Martín del Campo-Rodríguez, Helena Gómez-Adorno, Grigori Sidorov, and Ildar Batyrshin. "CIC-GIL Approach to Cross-domain Authorship Attribution". In: *Working Notes of CLEF* (2018) (cit. on p. 11).

[99] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 994–1003 (cit. on p. 61).

[100] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. "Unsupervised person re-identification: Clustering and fine-tuning". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.4 (2018), pp. 1–18 (cit. on pp. 61, 101).

[101] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets". In: *IEEE transactions on pattern analysis and machine intelligence* 41.3 (2018), pp. 523–536 (cit. on p. 51).

[102] Oren Halvani and Lukas Graner. "Cross-domain authorship attribution based on compression". In: *Working Notes of CLEF* (2018) (cit. on p. 29).

[103] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. "Masking: A new perspective of noisy supervision". In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 73).

[104] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. "Using trusted data to train deep networks on labels corrupted by severe noise". In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 73).

[105] Julian Hitschler, Esther Van Den Berg, and Ines Rehbein. "Authorship attribution with convolutional neural networks and POS-eliding". In: *Proceedings of the Workshop on Stylistic Variation (EMNLP 2017). September 8, 2017 Copenhagen, Denmark*. The Association for Computational Linguistics. 2018, pp. 53–28 (cit. on p. 27).

[106] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. "Adversarially occluded samples for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 5098–5107 (cit. on p. 61).

[107] Simon Jenni and Paolo Favaro. "Deep Bilevel Learning". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018 (cit. on p. 73).

[108] Yogiraj Kulkarni, Shubhangi Bodkhe, Amit Kamthe, and Archana Patil. "Automatic number plate recognition for motorcyclists riding without helmet". In: *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE. 2018, pp. 1–6 (cit. on p. 38).

[109] Rayson Laroca, Evair Severo, Luiz A Zanlorensi, Luiz S Oliveira, Gabriel Resende Gonçalves, William Robson Schwartz, and David Menotti. "A robust real-time automatic license plate recognition based on the YOLO detector". In: *2018 international joint conference on neural networks (ijcnn)*. IEEE. 2018, pp. 1–10 (cit. on p. 38).

[110] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. "Hyperband: A novel bandit-based approach to hyperparameter optimization". In: *Journal of Machine Learning Research* 18.185 (2018), pp. 1–52 (cit. on p. 97).

[111] Minxian Li, Xiatian Zhu, and Shaogang Gong. "Unsupervised Person Re-identification by Deep Learning Tracklet Association". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018 (cit. on p. 101).

[112] Wei Li, Xiatian Zhu, and Shaogang Gong. "Harmonious attention network for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2285–2294 (cit. on p. 101).

[113] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 67–83 (cit. on pp. 53–55).

[114] Sreenivas Mekala, Vishnu Vardan Bulusu, and Raghunadha Reddy. "A survey on authorship attribution approaches". In: *Int. J. Comput. Eng. Res.(IJCER)* 8.8 (2018) (cit. on p. 24).

[115] Jagadeesh Patchala and Raj Bhatnagar. "Authorship attribution by consensus among multiple features". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 2766–2777 (cit. on p. 24).

[116] Sergio Montazzolli Silva and Claudio Rosito Jung. "License plate detection and recognition in unconstrained scenarios". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 580–596 (cit. on p. 38).

[117] Efstathios Stamatatos. "Masking topic-related information to enhance authorship attribution". In: *Journal of the Association for Information Science and Technology* 69.3 (2018), pp. 461–473. DOI: https://doi.org/10.1002/asi.23968. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23968. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23968 (cit. on pp. 25, 26, 29).

[118] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. "Mancs: A multi-task attentional network with curriculum sampling for person re-identification". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 365–381 (cit. on p. 101).

[119] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. "Learning discriminative features with multiple granularities for person re-identification". In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 274–282 (cit. on p. 98).

[120] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. "Person transfer gan to bridge domain gap for person re-identification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 79–88 (cit. on pp. 57, 60).

[121] Jiwei Yang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. "Local convolutional neural networks for person re-identification". In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 1074–1082 (cit. on p. 98).

[122] Min Yang, Xiaojun Chen, Wenting Tu, Ziyu Lu, Jia Zhu, and Qiang Qu. "A topic drift model for authorship attribution". In: *Neurocomputing* 273 (2018), pp. 133–140 (cit. on p. 24).

[123] Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. "Deep learning from noisy image labels with quality embedding". In: *IEEE Transactions on Image Processing* 28.4 (2018), pp. 1909–1922 (cit. on p. 73).

[124] Yuan Yuan, Jian'an Zhang, and Qi Wang. "Bike-person re-identification: a benchmark and a comprehensive evaluation". In: *IEEE Access* 6 (2018), pp. 56059–56068 (cit. on p. 61).

[125] Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. "Syntax Encoding with Application in Authorship Attribution". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2742–2753. DOI: 10.18653/v1/D18-1294. URL: https://aclanthology.org/D18-1294 (cit. on pp. 24–26).

[126] Zhilu Zhang and Mert Sabuncu. "Generalized cross entropy loss for training deep neural networks with noisy labels". In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 73).

[127] Ehsan Amid, Manfred KK Warmuth, Rohan Anil, and Tomer Koren. "Robust bi-tempered logistic loss based on bregman divergences". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 73).

[128] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. "Character region awareness for text detection". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9365–9374 (cit. on p. 38).

[129] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. "Generalizing unmasking for short texts". In: *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019, pp. 654–659 (cit. on pp. 16, 26).

[130] Benedikt Boenninghoff, Robert M Nickel, Steffen Zeiler, and Dorothea Kolossa. "Similarity learning for authorship verification in social media". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 2457–2461 (cit. on pp. 12, 14).

[131] B. Bönninghoff, S. Hessler, D. Kolossa, and R. M. Nickel. "Explainable Authorship Verification in Social Media via Attention-based Similarity Learning". In: *2019 IEEE International Conference on Big Data (Big Data)*. Los Alamitos, CA, USA: IEEE Computer Society, 2019, pp. 36–45. DOI: 10.1109/BigData47090.2019.9005650. URL: https://doi.ieeecomputersociety.org/10.1109/BigData47090.2019.9005650 (cit. on pp. 27, 29).

[132] Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, and Jie Zhou. "Self-critical attention learning for person re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9637–9646 (cit. on p. 101).

[133] Rung-Ching Chen et al. "Automatic License Plate Recognition via sliding-window darknet-YOLO deep learning". In: *Image and Vision Computing* 87 (2019), pp. 47–56 (cit. on p. 38).

[134] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. "Learning resolution-invariant deep representations for person re-identification". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 8215–8222 (cit. on p. 61).

[135] Olga Fourkioti, Symeon Symeonidis, and Avi Arampatzis. "Language models and fusion for authorship attribution". In: *Information Processing & Management* 56.6 (2019), p. 102061 (cit. on p. 13).

[136] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. "Beyond human parts: Dual part-aligned representations for person re-identification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 3642–3651 (cit. on p. 101).

[137] Fereshteh Jafariakinabad and Kien A. Hua. "Style-Aware Neural Model with Application in Authorship Attribution". In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. 2019, pp. 325–328. DOI: 10.1109/ICMLA.2019.00061 (cit. on p. 27).

[138] Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien A Hua. "Syntactic recurrent neural network for authorship attribution". In: *arXiv preprint arXiv:1902.09723* (2019) (cit. on p. 11).

[139] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. "Set transformer: A framework for attention-based permutation-invariant neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 3744–3753 (cit. on p. 97).

[140] Younkwan Lee, Juhyun Lee, Hoyeon Ahn, and Moongu Jeon. "SNIDER: Single noisy image denoising and rectification for improving license plate recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 2019, pp. 0–0 (cit. on p. 38).

[141] Liu Leqi, Adarsh Prasad, and Pradeep K Ravikumar. "On human-aligned risk minimization". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 71, 73).

[142] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. "Global-local temporal representations for video person re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3958–3967 (cit. on p. 61).

[143] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. "A bottom-up clustering approach to unsupervised person re-identification". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 8738–8745 (cit. on p. 101).

[144] Hengyue Liu and Bir Bhanu. "Pose-guided R-CNN for jersey number recognition in sports". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0 (cit. on p. 39).

[145] Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. "Deep reinforcement active learning for human-in-the-loop person re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6122–6131 (cit. on p. 101).

[146] Yueming Lyu and Ivor W Tsang. "Curriculum loss: Robust learning and generalization against label corruption". In: *arXiv preprint arXiv:1905.10045* (2019) (cit. on p. 73).

[147] Jingke Meng, Sheng Wu, and Wei-Shi Zheng. "Weakly supervised person re-identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 760–769 (cit. on pp. 88, 89, 97, 101).

[148] Sauradip Nag, Raghavendra Ramachandra, Palaiahnakote Shivakumara, Umapada Pal, Tong Lu, and Mohan Kankanhalli. "CRNN based jersey-bib number/text recognition in sports and marathon images". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 1149–1156 (cit. on p. 39).

[149] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. "Human–machine partnership with artificial intelligence for chest radiograph diagnosis". In: *NPJ digital medicine* 2.1 (2019), pp. 1–10 (cit. on p. 71).

[150] Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. "TIRA Integrated Research Architecture". In: *Information Retrieval Evaluation in a Changing World*. Ed. by Nicola Ferro and Carol Peters. The Information Retrieval Series. Berlin Heidelberg New York: Springer, 2019. ISBN: 978-3-030-22948-1. DOI: 10.1007/978-3-030-22948-1_5 (cit. on p. 16).

[151] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084 (cit. on p. 13).

[152] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084 (cit. on p. 27).

[153] Yanyao Shen and Sujay Sanghavi. "Learning with bad training data via iterative trimmed loss minimization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 5739–5748 (cit. on p. 73).

[154] Yanyao Shen and Sujay Sanghavi. "Learning with Bad Training Data via Iterative Trimmed Loss Minimization". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 5739–5748. URL: https://proceedings.mlr.press/v97/shen19e.html (cit. on p. 73).

[155] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.9 (2019), pp. 2035–2048. DOI: 10.1109/TPAMI.2018.2848939 (cit. on p. 38).

[156] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. "Meta-weight-net: Learning an explicit mapping for sample weighting". In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 73).

[157] Leslie N Smith and Nicholay Topin. "Super-convergence: Very fast training of neural networks using large learning rates". In: *Artificial intelligence and machine learning for multi-domain operations applications*. Vol. 11006. SPIE. 2019, pp. 369–386 (cit. on p. 76).

[158] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. "Generalizable person re-identification by domain-invariant mapping network". In: *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2019, pp. 719–728 (cit. on p. 101).

[159] PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. "Multiple instance learning for histopathological breast cancer image classification". In: *Expert Systems with Applications* 117 (2019), pp. 103–111 (cit. on pp. 88, 101).

[160] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. "C-mil: Continuation multiple instance learning for weakly supervised object detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2199–2208 (cit. on p. 101).

[161] Ruishuang Wang, Zhao Li, Jian Cao, Tong Chen, and Lei Wang. "Convolutional Recurrent Neural Networks for Text Classification". In: *2019 International Joint Conference on Neural Networks (IJCNN)*. 2019, pp. 1–6. DOI: 10.1109/IJCNN.2019.8852406 (cit. on p. 38).

[162] Qize Yang, Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. "Patch-based discriminative feature learning for unsupervised person re-identification". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 3633–3642 (cit. on p. 101).

[163] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. "Unsupervised Person Re-Identification by Soft Multilabel Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019 (cit. on p. 101).

[164] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. "Icdar 2019 robust reading challenge on read-

ing chinese text on signboard". In: *2019 international conference on document analysis and recognition (ICDAR)*. IEEE. 2019, pp. 1577–1581 (cit. on p. 44).

[165] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. "Omni-scale feature learning for person re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 3702–3712 (cit. on pp. 52, 55–57, 60).

[166] Sanping Zhou, Fei Wang, Zeyi Huang, and Jinjun Wang. "Discriminative feature learning with consistent attention regularization for person re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 8040–8049 (cit. on p. 101).

[167] N Palanivel Ap, T Vigneshwaran, M Sriv Arappradhan, and R Madhanraj. "Automatic number plate detection in vehicles using faster R-CNN". In: *2020 International conference on system, computation, automation and networking (ICSCAN)*. IEEE. 2020, pp. 1–6 (cit. on p. 38).

[168] Georgios Barlas and Efstathios Stamatatos. "Cross-domain authorship attribution using pre-trained language models". In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer. 2020, pp. 255–266 (cit. on pp. 13, 25, 27, 29).

[169] Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. "The importance of suppressing domain style in authorship analysis". In: *arXiv preprint arXiv:2005.14714* (2020) (cit. on pp. 25, 26, 29).

[170] Benedikt Boenninghoff, Julian Rupp, Robert M Nickel, and Dorothea Kolossa. "Deep bayes factor scoring for authorship verification". In: *arXiv preprint arXiv:2008.10105* (2020) (cit. on pp. 11, 25, 27).

[171] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607 (cit. on p. 98).

[172] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. "Improved baselines with momentum contrastive learning". In: *arXiv preprint arXiv:2003.04297* (2020) (cit. on p. 98).

[173] Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. "Adaptive sampling for stochastic risk-averse learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1036–1047 (cit. on p. 73).

[174] Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. "CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 6903–6915. DOI: 10.18653/v1/2020.coling-main.609. URL: https://aclanthology.org/2020.coling-main.609 (cit. on p. 27).

[175] Maël Fabien, Esa ú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. "BertAA: BERT fine-tuning for Authorship Attribution". In: *Proceedings of the 17th International Conference on Natural Language Processing*. CONF. ACL. 2020 (cit. on pp. 13, 22, 24, 25, 28, 29).

[176] Martin Gerlach and Francesc Font-Clos. "A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics". In: *Entropy* 22.1 (2020), p. 126 (cit. on p. 25).

[177] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. "Bootstrap your own latent-a new approach to self-supervised learning". In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284 (cit. on p. 98).

[178] Fereshteh Jafariakinabad and Kien A Hua. "A Self-supervised Representation Learning of Sentence Structure for Authorship Attribution". In: *arXiv preprint arXiv:2010.06786* (2020) (cit. on p. 11).

[179] Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. "Overview of the Cross-Domain Authorship Verification Task at PAN 2020". In: *CLEF*. 2020 (cit. on p. 15).

[180] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale". In: *IJCV* (2020) (cit. on p. 43).

[181] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. "Unsupervised person re-identification via softened similarity learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 3390–3399 (cit. on p. 101).

[182] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. "Deep learning for generic object detection: A survey". In: *International journal of computer vision* 128 (2020), pp. 261–318 (cit. on p. 1).

[183] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. "Abcnet: Real-time scene text spotting with adaptive bezier-curve network". In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9809–9818 (cit. on p. 44).

[184] Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. "Towards improved model design for authorship identification: A survey on writing style understanding". In: *arXiv preprint arXiv:2009.14445* (2020) (cit. on p. 24).

[185] Juanita Ordoñez, Rafael Rivera Soto, and Barry Y Chen. "Will longformers PAN out for authorship verification". In: *Working Notes of CLEF* (2020) (cit. on p. 26).

[186] Adrian Penate-Sanchez, David Freire-Obregón, Adrián Lorenzo-Melián, Javier Lorenzo-Navarro, and Modesto Castrillón-Santana. "TGC20ReId: A dataset for sport event re-identification in the wild". In: *Pattern Recognition Letters* 138 (2020), pp. 355–361. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2020.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0167865520303019 (cit. on p. 61).

[187] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. "CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV". In: *NPJ digital medicine* 3.1 (2020), pp. 1–8 (cit. on p. 71).

[188] Jithmi Shashirangana, Heshan Padmasiri, Dulani Meedeniya, and Charith Perera. "Automated license plate recognition: a survey on methods and techniques". In: *IEEE Access* 9 (2020), pp. 11203–11225 (cit. on p. 36).

[189] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. "Human–computer collaboration for skin cancer recognition". In: *Nature Medicine* 26.8 (2020), pp. 1229–1234 (cit. on p. 71).

[190] Dongkai Wang and Shiliang Zhang. "Unsupervised Person Re-Identification via Multi-Label Classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 (cit. on p. 101).

[191] Guangrun Wang, Guangcong Wang, Xujie Zhang, Jianhuang Lai, Zhengtao Yu, and Liang Lin. "Weakly supervised person re-id: Differentiable graphical learning and a new benchmark". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.5 (2020), pp. 2142–2156 (cit. on pp. 88, 98, 101).

[192] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. "Generalizing from a few examples: A survey on few-shot learning". In: *ACM computing surveys (csur)* 53.3 (2020), pp. 1–34 (cit. on p. 1).

[193] Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. "Dual t: Reducing estimation error for transition matrix in label-noise learning". In: *Advances in neural information processing systems* 33 (2020), pp. 7260–7271 (cit. on p. 73).

[194] Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. "The Topic Confusion Task: A Novel Evaluation Scenario for Authorship Attribution". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 4242–4256. DOI: 10.18653/v1/2021.findings-emnlp.359. URL: https://aclanthology.org/2021.findings-emnlp.359 (cit. on pp. 22, 25, 28, 29).

[195] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. "DocFormer: End-to-End Transformer for Document Understanding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 993–1003 (cit. on p. 36).

[196] Rowel Atienza. "Vision Transformer for Fast and Efficient Scene Text Recognition". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by Josep Lladós, Daniel Lopresti, and Seiichi Uchida. Cham: Springer International Publishing, 2021, pp. 319–334. ISBN: 978-3-030-86549-8 (cit. on p. 38).

[197] Janek Bevendorff, BERTa Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. "Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection". In: *12th International Conference of the CLEF Association (CLEF 2021)* (Bucharest, Romania). Springer, 2021 (cit. on p. 12).

[198] Benedikt T. Bönninghoff, Robert M. Nickel, and Dorothea Kolossa. "O2D2: Out-Of-Distribution Detector to Capture Undecidable Trials in Authorship Verification". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi. Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 1846–1857. URL: http://ceur-ws.org/Vol-2936/paper-158.pdf (cit. on pp. 25, 27).

[199] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. "Explainable machine learning in credit risk management". In: *Computational Economics* 57.1 (2021), pp. 203–216 (cit. on p. 71).

[200] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. "Text Recognition in the Wild: A Survey". In: *ACM Comput. Surv.* 54.2 (2021). ISSN: 0360-0300. DOI: 10.1145/3440756. URL: https://doi.org/10.1145/3440756 (cit. on pp. 36, 38).

[201] John C Duchi and Hongseok Namkoong. "Learning models with uniform performance via distributionally robust optimization". In: *The Annals of Statistics* 49.3 (2021), pp. 1378–1406 (cit. on p. 73).

[202] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. "A brief review of domain adaptation". In: *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020* (2021), pp. 877–894 (cit. on p. 1).

[203] Augusto Figueiredo, Johnata Brayan, Renan Oliveira Reis, Raphael Prates, and William Robson Schwartz. "More: a large-scale motorcycle re-identification dataset". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 4034–4043 (cit. on p. 61).

[204] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. "Unsupervised Pre-Training for Person Re-Identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 14750–14759 (cit. on p. 101).

[205] Romain Futrzynski. "Author classification as pre-training for pairwise authorship verification". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi. Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 1945–1952. URL: http://ceur-ws.org/Vol-2936/paper-168.pdf (cit. on pp. 22, 25, 28).

[206] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. "Yolox: Exceeding yolo series in 2021". In: *arXiv preprint arXiv:2107.08430* (2021) (cit. on p. 55).

[207] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. "Transreid: Transformer-based object re-identification". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 15013–15022 (cit. on pp. 60, 101).

[208] Pablo Hernández-Carrascosa, Adrian Penate-Sanchez, Javier Lorenzo-Navarro, David Freire-Obregón, and Modesto Castrillón-Santana. "TGCRBNW: A Dataset for Runner Bib Number Detection (and Recognition) in the Wild". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 9445–9451. DOI: 10.1109/ICPR48806.2021.9412220 (cit. on p. 39).

[209] Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Benno Stein, and Martin Potthast. "Overview of the Cross-Domain Authorship Verification Task at PAN 2021". In: *CLEF (Working Notes)*. 2021, pp. 1743–1759. URL: http://ceur-ws.org/Vol-2936/paper-147.pdf (cit. on pp. 22, 24, 25, 27, 29).

[210] Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Janek Bevendorff, Martin Potthast, and Benno Stein. "Overview of the Authorship Verification Task at PAN 2021". In: *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2021 (cit. on pp. 15, 16).

[211] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. "Open images v5 text annotation and yet another mask text spotter". In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 379–389 (cit. on pp. 42, 43).

[212] Kseniya Vladimirovna Lagutina. "Comparison of style features for the authorship verification of literary texts". In: *Modeling and analysis of information systems* 28.3 (2021), pp. 250–259 (cit. on p. 25).

[213] Rayson Laroca, Luiz A Zanlorensi, Gabriel R Gonçalves, Eduardo Todt, William Robson Schwartz, and David Menotti. "An efficient and layout-independent automatic license plate recognition system based on the YOLO detector". In: *IET Intelligent Transport Systems* 15.4 (2021), pp. 483–503 (cit. on p. 38).

[214] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. "Tilted Empirical Risk Minimization". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=K5YasWXZT3O (cit. on pp. 72, 73).

[215] Hao Luo, Pichao Wang, Yi Xu, Feng Ding, Yanxin Zhou, Fan Wang, Hao Li, and Rong Jin. "Self-supervised pre-training for transformer-based person re-identification". In: *arXiv preprint arXiv:2111.12084* (2021) (cit. on p. 101).

[216] Andrei Manolache, Florin Brad, Elena Burceanu, Antonio Barbalau, Radu Ionescu, and Marius Popescu. "Transferring BERT-like Transformers' Knowledge for Authorship Verification". In: *arXiv preprint arXiv:2112.05125* (2021) (cit. on pp. 25, 27).

[217] Jingke Meng, Wei-Shi Zheng, Jian-Huang Lai, and Liang Wang. "Deep graph metric learning for weakly supervised person re-identification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2021), pp. 6074–6093 (cit. on p. 88).

[218] Benjamin Murauer and Günther Specht. "Developing a Benchmark for Reducing Data Bias in Authorship Attribution". In: *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 179–188. DOI: 10.18653/v1/2021.eval4nlp-1.18. URL: https://aclanthology.org/2021.eval4nlp-1.18 (cit. on pp. 22, 24, 25, 29).

[219] Zeyang Peng, Leilei Kong, Zhijie Zhang, Zhongyuan Han, and Xu Sun. "Encoding Text Information By Pre-trained Model For Authorship Verification". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi. Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 2103–2107. URL: http://ceur-ws.org/Vol-2936/paper-186.pdf (cit. on pp. 22, 25, 28).

[220] Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. "Learning Universal Authorship Representations". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 913–919. DOI: 10.18653/v1/2021.emnlp-main.70. URL: https://aclanthology.org/2021.emnlp-main.70 (cit. on pp. 25, 27).

[221] Chakaveh Saedi and Mark Dras. "Siamese networks for large-scale author identification". In: *Computer Speech & Language* 70 (2021), p. 101241. ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2021.101241. URL: https://www.sciencedirect.com/science/article/pii/S0885230821000486 (cit. on pp. 25, 26).

[222] S Sanjana, S Sanjana, VR Shriya, Gururaj Vaishnavi, and K Ashwini. "A review on various methodologies used for vehicle classification, helmet detection and number plate recognition". In: *Evolutionary Intelligence* 14.2 (2021), pp. 979–987 (cit. on p. 38).

[223] Richard Sinnott and Zijian Wang. "Linking User Accounts across Social Media Platforms". In: *2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies (BDCAT'21)*. 2021, pp. 18–27 (cit. on p. 21).

[224] Jacob Tyo, Bhuwan Dhingra, and Zachary Lipton. "Siamese Bert for Authorship Verification". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi. Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 2169–2177. URL: http://ceur-ws.org/Vol-2936/paper-193.pdf (cit. on pp. 22, 25, 28, 29).

[225] Kanav Vats, Mehrnaz Fani, David A Clausi, and John Zelek. "Multi-task learning for jersey number recognition in ice hockey". In: *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*. 2021, pp. 11–15 (cit. on p. 39).

[226] Guangrun Wang, Keze Wang, Guangcong Wang, Philip HS Torr, and Liang Lin. "Solving inefficiency of self-supervised representation learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 9505–9515 (cit. on p. 98).

[227] Xueping Wang, Min Liu, Dripta S Raychaudhuri, Sujoy Paul, Yaonan Wang, and Amit K Roy-Chowdhury. "Learning person re-identification models from videos with weak supervision". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 3017–3028 (cit. on pp. 88, 101).

[228] Janith Weerasinghe, Rhia Singh, and Rachel Greenstadt. "Feature Vector Difference based Authorship Verification for Open-World Settings". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*. Ed. by Guglielmo Faggioli, Nicola Ferro, Alexis Joly, Maria Maistro, and Florina Piroi. Vol. 2936. CEUR Workshop Proceedings. CEUR-WS.org, 2021, pp. 2201–2207. URL: http://ceur-ws.org/Vol-2936/paper-197.pdf (cit. on pp. 25, 26, 29).

[229] Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. "Estimating instance-dependent label-noise transition matrix using dnns". In: *arXiv preprint arXiv:2105.13001* (2021) (cit. on p. 73).

[230] Mang Ye, He Li, Bo Du, Jianbing Shen, Ling Shao, and Steven CH Hoi. "Collaborative refining for person re-identification with label noise". In: *IEEE Transactions on Image Processing* 31 (2021), pp. 379–391 (cit. on pp. 89, 97, 101).

[231] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. "Deep learning for person re-identification: A survey and outlook". In: *IEEE transactions on pattern analysis and machine intelligence* 44.6 (2021), pp. 2872–2893 (cit. on pp. 51, 60, 96, 101).

[232] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. "Multiple instance active learning for object detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5330–5339 (cit. on p. 101).

[233] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. "Weakly supervised text-based person re-identification". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11395–11404 (cit. on pp. 88, 101).

[234] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. "Weakly supervised contrastive learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10042–10051 (cit. on pp. 88, 101).

[235] Darwin Bautista and Rowel Atienza. "Scene Text Recognition with Permuted Autoregressive Sequence Models". In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022, pp. 178–196. DOI: 10.1007/978-3-031-19815-1_11. URL: https://doi.org/10.1007/978-3-031-19815-1_11 (cit. on p. 38).

[236] Divya Bhargavi, Erika Pelaez Coyotl, and Sia Gholami. "Knock, knock. Who's there?–Identifying football player jersey numbers with synthetic data". In: *arXiv preprint arXiv:2203.00734* (2022) (cit. on p. 39).

[237] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. "Part-based pseudo label refinement for unsupervised person re-identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 7308–7318 (cit. on p. 101).

[238] Charles-Alexandre Diop, Baptiste Pelloux, Xinrui Yu, Won-Jae Yi, and Jafar Saniie. "Soccer Player Recognition using Artificial Intelligence and Computer Vision". In: *2022 IEEE International Conference on Electro Information Technology (eIT)*. IEEE. 2022, pp. 477–481 (cit. on p. 39).

[239] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. "Distributionally robust losses for latent covariate mixtures". In: *Operations Research* (2022) (cit. on p. 73).

[240] Daniel Embarcadero-Ruiz, Helena Gómez-Adorno, Alberto Embarcadero-Ruiz, and Gerardo Sierra. "Graph-Based Siamese Network for Authorship Verification". In: *Mathematics* 10.2 (2022). ISSN: 2227-7390. DOI: 10.3390/math10020277. URL: https://www.mdpi.com/2227-7390/10/2/277 (cit. on pp. 25, 26).

[241] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. "Large-Scale Pre-Training for Person Re-Identification With Noisy Labels". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 2476–2486 (cit. on p. 101).

[242] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, et al. "SoccerNet 2022 challenges results". In: *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. 2022, pp. 75–86 (cit. on p. 61).

[243] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. "Swintextspotter: Scene text spotting via better synergy between text detection and text recognition". In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4593–4603 (cit. on pp. 43, 44).

[244] Fereshteh Jafariakinabad and Kien A. Hua. "A Self-Supervised Representation Learning of Sentence Structure for Authorship Attribution". In: *ACM Trans. Knowl. Discov. Data* 16.4 (2022). ISSN: 1556-4681. DOI: 10.1145/3491203. URL: https://doi.org/10.1145/3491203 (cit. on pp. 24, 25, 27, 31).

[245] Jiaze Li and Bin Liu. "Rider Re-identification Based on Pyramid Attention". In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer. 2022, pp. 81–93 (cit. on p. 61).

[246] Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. "VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web". In: *arXiv preprint arXiv:2207.03477* (2022) (cit. on p. 28).

[247] Huy Che Quang, Tung Do Thanh, and Cuong Truong Van. "Character Time-series Matching For Robust License Plate Recognition". In: *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE. 2022, pp. 1–6 (cit. on p. 38).

[248] Pushkar Sathe, Aditi Rao, Aditya Singh, Ritika Nair, and Abhilash Poojary. "Helmet Detection And Number Plate Recognition Using Deep Learning". In: *2022 IEEE Region 10 Symposium (TENSYMP)*. IEEE. 2022, pp. 1–6 (cit. on p. 38).

[249] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. "Deepsportradar-v1: Computer vision dataset for sports understanding with high quality annotations". In: *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*. 2022, pp. 1–8 (cit. on p. 61).

[250] William Wong, Audrey Huang, Liu Leqi, Kamyar Azizzadenesheli, and Zachary C Lipton. "RiskyZoo: A Library for Risk-Sensitive Supervised Learning". In: (2022) (cit. on pp. xx, 71, 72).

[251] Zizheng Yang, Xin Jin, Kecheng Zheng, and Feng Zhao. "Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14298–14307 (cit. on p. 101).

[252] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. "Pass: Part-aware self-supervised pre-training for person re-identification". In: *European Conference on Computer Vision*. Springer. 2022, pp. 198–214 (cit. on p. 101).

[253] Ragd Alhejaily, Rahaf Alhejaily, Mai Almdahrsh, Shareefah Alessa, and Saleh Albelwi. "Automatic Team Assignment and Jersey Number Recognition in Football Videos". In: *INTELLIGENT AUTOMATION AND SOFT COMPUTING* 36.3 (2023), pp. 2669–2684 (cit. on p. 39).

[254] Saurabh Garg, Amrith Setlur, Zachary Chase Lipton, Sivaraman Balakrishnan, Virginia Smith, and Aditi Raghunathan. "Complementary Benefits of Contrastive Learning and Self-Training Under Distribution Shift". In: *arXiv preprint arXiv:2312.03318* (2023) (cit. on pp. 1, 88, 101).

[255] Min Liu, Yuan Bian, Qing Liu, Xueping Wang, and Yaonan Wang. "Weakly Supervised Tracklet Association Learning with Video Labels for Person Re-identification". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023) (cit. on pp. 88, 98, 99, 101).

[256] Jun Shu, Xiang Yuan, Deyu Meng, and Zongben Xu. "Cmw-net: Learning a class-aware sample weighting mapping for robust deep learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023) (cit. on p. 73).

[257] Jacob Tyo, Motolani Olarinre, Youngseog Chung, and Zachary C Lipton. "MUDD: A New Re-Identification Dataset with Efficient Annotation for Off-Road Racers in Extreme Conditions". In: *arXiv preprint arXiv:2311.08488* (2023) (cit. on pp. 83, 91).

[258] Suncheng Xiang, Dahong Qian, Jingsheng Gao, Zirui Zhang, Ting Liu, and Yuzhuo Fu. "Rethinking person re-identification via semantic-based pretraining". In: *ACM Transactions on Multimedia Computing, Communications and Applications* 20.3 (2023), pp. 1–17 (cit. on p. 101).

[259] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. "DeepSolo: Let Transformer Decoder With Explicit Points Solo for Text Spotting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 19348–19357 (cit. on pp. 38, 44).

[260] Jongmin Yu, Hyeontaek Oh, Minkyung Kim, and Junsik Kim. "Weakly supervised contrastive learning for unsupervised vehicle reidentification". In: *IEEE transactions on neural networks and learning systems* (2023) (cit. on p. 101).

[261] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. *CLIP4STR: A Simple Baseline for Scene Text Recognition with Pre-trained Vision-Language Model*. 2023. arXiv: 2305.14014 [cs.CV] (cit. on p. 38).