Spontaneously Emerging Object Part Segmentation

Yijie Wang
Machine Learning Department
Carnegie Mellon University
yijiewang@cmu.edu

Katerina Fragkiadaki
Machine Learning Department
Carnegie Mellon University
katef@cs.cmu.edu

Abstract

Object part is a fine-grained understanding of objects. It is also a generic model of objects in the same category. People recognize object subparts naturally, and meaningful object part segmentation exists for almost all natural objects. However, not many research works are conducted on the subpart level of generic object types, probably due to the lack of available data sources. In this project, we explore the possibility of forming object part segmentation without direct supervision. We apply deep learning with specifically designed loss on the ShapeNet dataset, and demonstrate a plausible system that can discover consistent and understandable object part segmentation spontaneously.

1 Introduction

What natural object does not have subparts? Most natural objects contain meaningful subparts. Object part is a fine-grained understanding of the objects, and the segmentation of object parts can depend on the function, material, texture or geometric shape of subparts. The subparts of an object category constitute a generic model of that category. Subparts of cars can be wheel, door, hood, roof or trunk because most cars have those parts and the relative location, geometric shape, function and material of those subparts are usually consistent.

However, there are very few research works on the subparts of generic objects, probably due to the lack of available labeled dataset. Only extensive research of human subparts is present, including human pose estimation and human facial landmark localization. Human pose estimation typically involves the localization of joints. Facial landmark localization is the task of identifying facial subparts. But most works in this field are supervised learning of clearly defined subparts.

Given the scarcity of generic object part datasets, we propose to create a deep learning system that discovers object part without explicit supervision, using 3D consistency loss. This system requires 3D models of objects which can be rendered in different poses, and can proposes reasonable and consistent object part segmentation on the rendered image.

2 Related Work

There are extensive researches on the localization of human subparts, including pose estimation and facial landmark localization. Human pose estimation typically involves the localization of joints, and there are many recent work on this topic. [5] is a exemplary work that estimates human pose using depth images. [6] is a more recent project that greatly improves human pose estimation using deep learning. Facial landmark localization is the task of identifying facial subparts, and has a much longer history in academia. Note that landmarks annotations are typically much finer than face parts in common-sense. [4] is a famous dataset of facial landmark annotation that leads to the development of many related works. [10], [9] and [8] are all examples of works on facial landmark localization.

Data Analysis Project, Master of Science in Machine Learning, Carnegie Mellon University

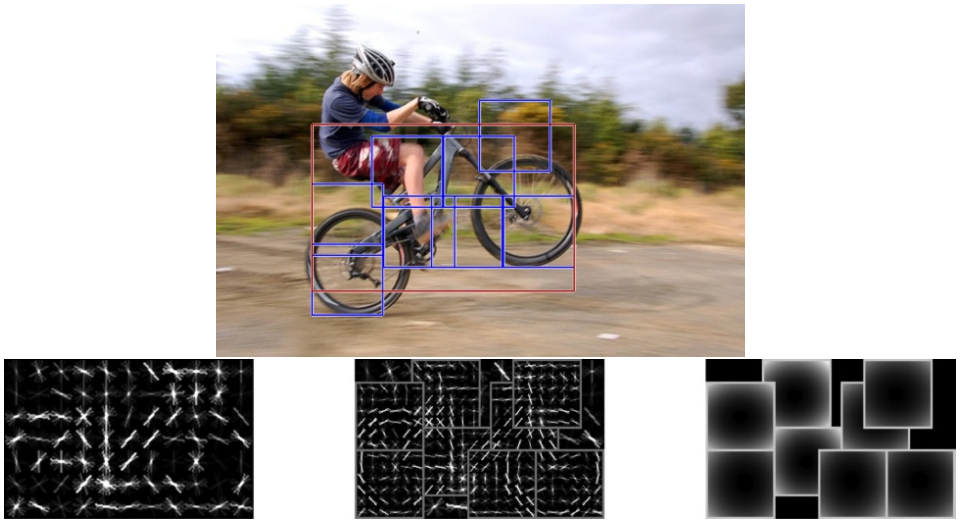

Figure 1: Learned template of a bicycle using deformable part model [2].

Deformable part model [2] was state-of-the-art object localization method before deep learning. It mostly involves a two-level hierarchy in which the bottom level represents parts of objects with gradient-based template and the top level captures the relative locations of parts that forms a more generic template of objects in a certain category. Figure 1 illustrates the learned template of a bicycle from [2]. However, because the part model is only based on histogram of oriented gradients, the learned part template in deformable part models is not performing perfectly and does not correspond well to part segmentation from common sense.

Recent development in deep learning leads to a much finer representation of pixels. [7] investigates the problem of creating a generic vector embedding for each pixels in a image. The embedding is trained with the task of predicting relative camera pose and pixel matching, and is demonstrated to generalize well to other tasks including scene layout, object pose and surface norm prediction. However, the vector embedding for each pixel in the image is not understandable for human, is not discrete and does not reveal any grouping or clustering that could be interpreted as object parts.

3 Dataset

Figure 2: Different car models in ShapeNet [1].

In this project, we use ShapeNet [1] as our primary dataset. ShapeNet is a collection of over 50000 3D models of objects in 55 common object categories. All 3D models are colorful, normalized in scale and oriented based on common sense. In this project, we focus on the car category because it is the largest category in ShapeNet in terms of model size, and because cars has many commonly agreed subparts. ShapeNet contains 7497 3D car models.

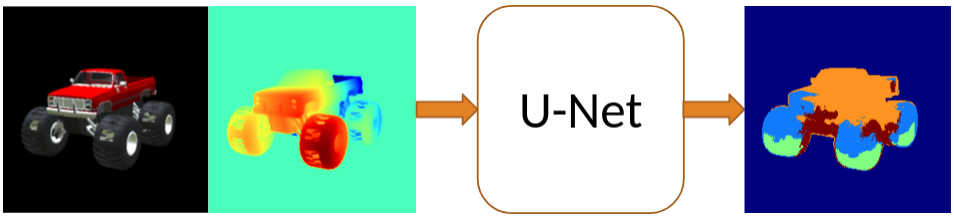

Figure 3: The essential task of part segmentation.

4 Method

Illustrated in Figure 3, the essential task of this project is part segmentation, in which an image, its depth map and the object mask is fed into a neural network, producing classification result of each pixel of the object. The network structure we use is identical to U-Net [3]. However, unlike in [3] where the class of each pixel is known and supervised, in our project the class of each pixel is unknown and the meaning of each class is undefined. Therefore, weak supervision is required to produce meaningful part segmentation result.

4.1 3D Consistency as Weak Supervision

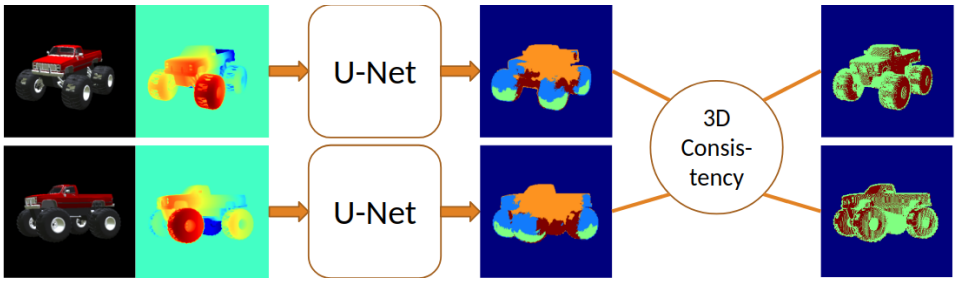

Figure 4: The concept of 3D consistency. The rightmost two images illustrates the pairs of matched pixels in the two views, where red pixels means matched pixels and green ones are unmatched pixels.

What properties should object parts have? Object parts should be consistent, meaning that pose change should not affect the segmentation of object parts. That leads to 3D consistency loss, which we use as weak supervision. Object parts should also be connected and concentrated, which is realized by the inductive bias of the U-Net architecture.

Specifically, 3D consistency means that corresponding pixels in two images of the same object should have identical classification. In order to enforce 3D consistency during training, we render two images of a randomly selected 3D model from two views and record the correspondence between pixels and 3D points on the model. Pixels from the two images are considered matched pixel pairs if their corresponding 3D point on the model is less than 0.01 apart.

We define a distance metric D_G of two discrete probability distribution inspired by the Gini impurity as:

$$D_G(P, Q) = 1 - \sum_c P(x = c)Q(x = c)$$

and given the set of matching pixels $M = \{(p_i, q_i)\}_{i=1}^N$, the 3D consistency loss is defined as:

$$L_{3D}(x^{(1)}, x^{(2)}) = \frac{1}{N} \sum_{(p_i,q_i)\in M} D_G(x_{p_i}^{(1)}, x_{q_i}^{(2)})$$

where $x^{(1)}$, $x^{(2)}$ are the classification result of the two images, and $x_{p_i}^{(1)}$ represents the probability distribution of classes at location p_i in $x^{(1)}$.

4.2 Class Diversity

With only 3D consistency as weak supervision, the model always degrades to a trivial solution that classifies all pixels into a single class. In that case the matching pixels always have identical classification. Therefore, we need to encourage the diversity of classification result. Specifically, let the object mask as a set of pixels be $O = \{p_i\}_{i=1}^{M}$, and the diversity loss is:

$$L_{div}(x) = -\mathbb{E}_{p_1,p_2 \sim O} D_G(x_{p_1}, x_{p_2})$$

where x is the classification result of an image, p_1, p_2 are two locations randomly drawn from the object mask, x_{p_1} is the class probability distribution at location p_1, and D_G is the distance metric specified in the previous subsection.

4.3 Hierarchical Classification

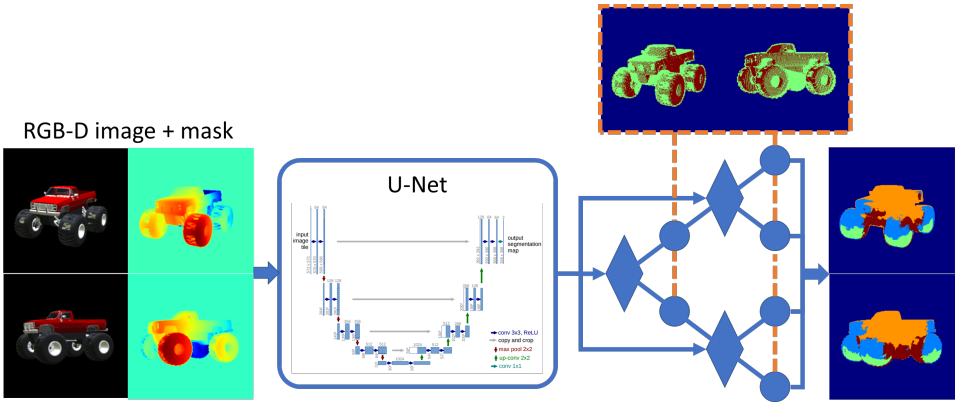

Figure 5: Illustration of hierarchical classification.

Unfortunately the diversity loss specified in the previous subsection could not perfectly improve the diversity of classification result alone. No matter how many possible classes the network architecture allows, the model always collapse into outputting only two classes. Tuning hyperparameters does not help. But obviously objects could have more than two parts.

In order to alleviate this problem, we incorporate hierarchical classification into the system. At each node of the hierarchical classification tree, only a binary classification is performed. By employing both 3D consistency loss and diversity loss, we can produce reasonable binary classification result at each node. By organizing binary classification into a hierarchical classification tree, we obtain reasonably diverse classification result. The hierarchical classification is illustrated in Figure 5.

5 Results

The U-Net is trained with 3D consistency loss and diversity loss on the hierarchical classification scheme using the momentum method. No pretrained parameters are used in the network. For the best visualization result, we choose to classify the object into four classes, which is realized using a classification tree of size 3 as illustrated in Figure 5. In order to generate training inputs, we randomly choose from the 3D car models and render a pair of images from different views. The views are limited so that the viewing point is above the "ground" because chassis is sometimes poorly modeled, or not modeled at at, in ShapeNet. The views are limited to be at most 30 degrees apart in yaw or pitch, so that there are enough matching pixels for 3D consistency.

Figure 6 demonstrates the object part segmentation results of a sedan from different views. The third and sixth columns are the object part segmentation results. The sedan is segmented into four parts. The yellow part is roof and windows, the brown part doors, the blue part hood and trunk, and the green part wheel and bumpers. According to the visualization, the parts are connected, and the 3D consistency is well maintained. The diversity of classification is well enforced, and the size of each part is balanced. On the first row where the views are front and rear views, the brown

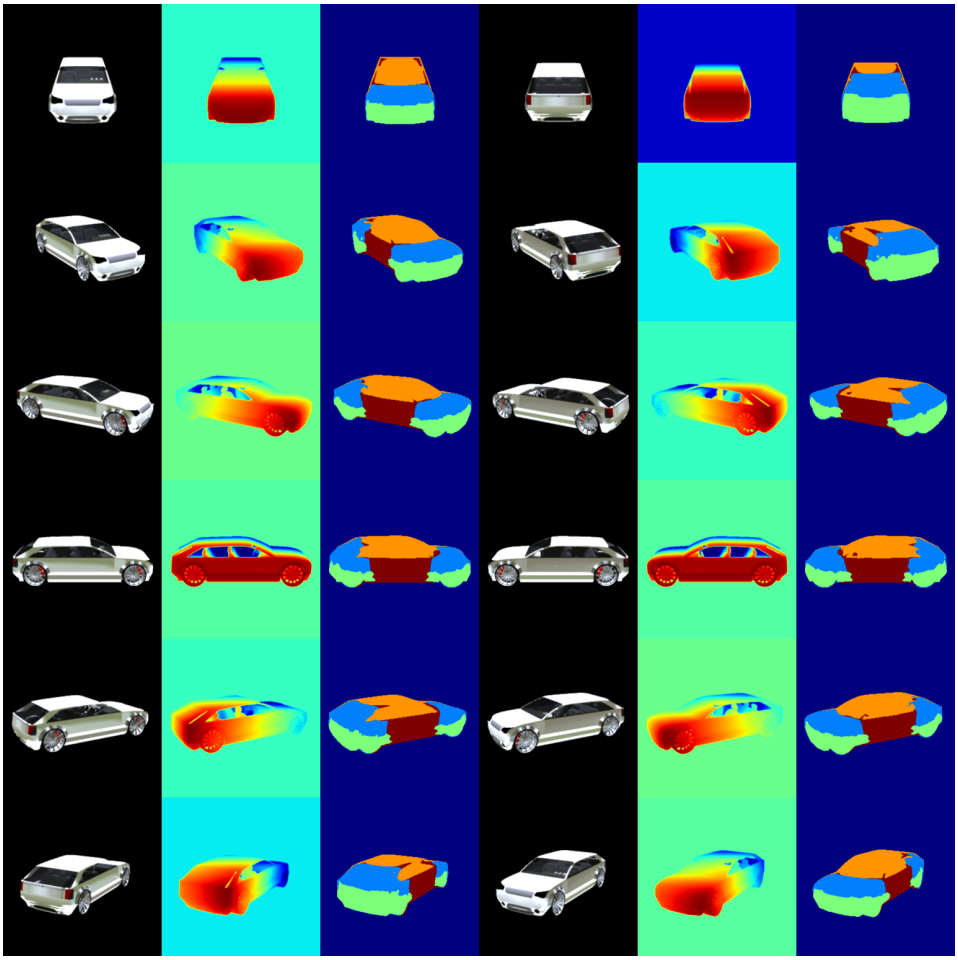

Figure 6: Visualization of object part segmentation results of a sedan which is common in the dataset. The columns are in the order of input image, depth and classification result.

part, representing doors, disappears as expected, which suggests that the diversity loss can tolerate temporarily hidden parts.

Figure 7 visualizes the object part segmentation results of a truck, which is much less common in the ShapeNet dataset than sedan. Interestingly, the brown part is still doors, the green part still contains wheel and bumpers, the blue part still contains hood, and the yellow part still contains windows. This suggests that the model is indeed learning a somewhat generic model of cars, so that the contents of each object part is relatively stable and understandable. However, in this case, because trucks are rare in ShapeNet, the model could not produce stable part segmentation for the rear vertical surface of this truck. The segmentation result from the rear view on top right is a failure example, where the 3D consistency is apparently broken.

6 Conclusion, Limitation and Future Work

In this project, we demonstrate a neural network system that can produce reasonable object part segmentation without explicit supervision. The weak supervision, namely 3D consistency loss, is able to produce object part segmentation spontaneously, when the class diversity is maintained by the diversity loss and hierarchical classification scheme.

However, this project is limited in several ways. First, the definition of object part is unclear. 3D consistency is definitely one property of object part segmentation, but any segmentation that is 3D consistent may not be a good object part segmentation. Object part could involve material,

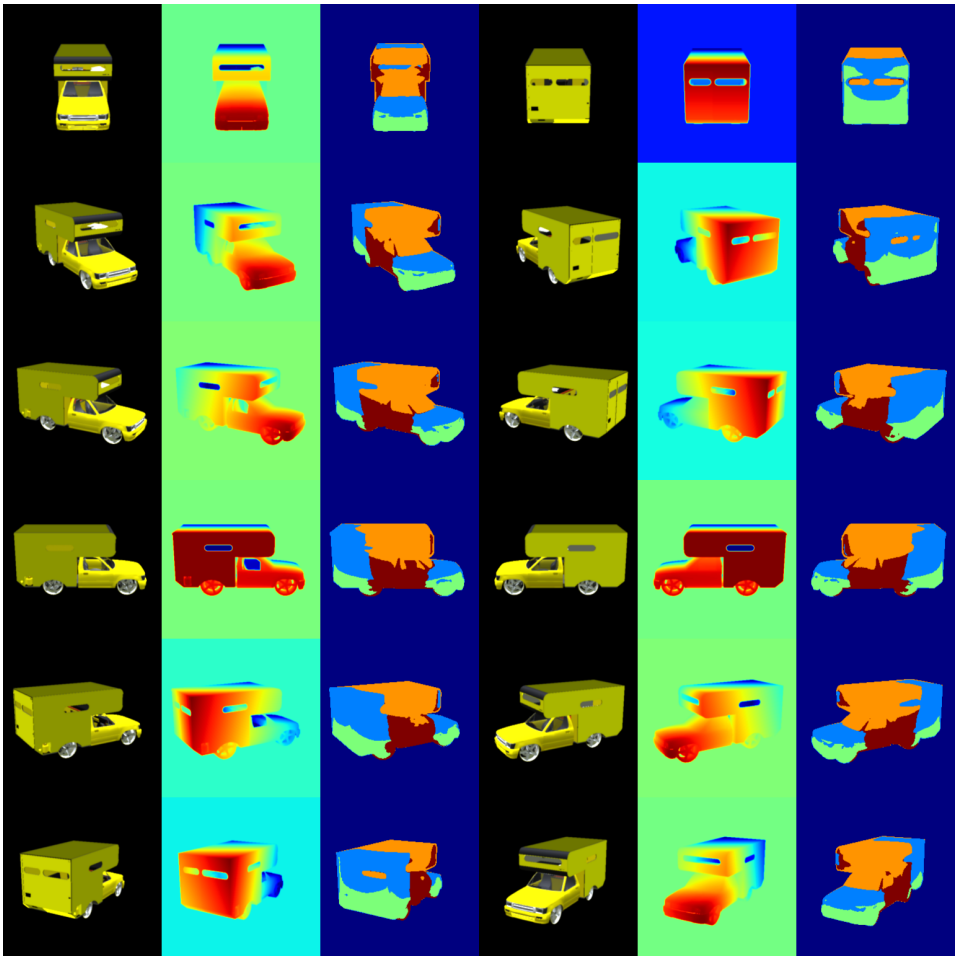

Figure 7: Visualization of object part segmentation results of a truck which is less common in the dataset. The columns are in the order of input image, depth and classification result.

functionality or geometric shapes, which is not reflected in this system. Second, evaluation of the results is not sufficient, because we have not implemented a quantitative way to measure its quality. 3D consistency is not a fair metric to compare our method to other unsupervised methods, because our method explicitly optimizes for 3D consistency.

We are aware of a very recent competition in ICCV 2017 Workshop on Learning to See from 3D Data, called Large-Scale 3D Shape Reconstruction and Segmentation from ShapeNet Core55, that includes a dataset of part segmentation of 3D point clouds sampled from models in certain ShapeNet categories. The task of this competition is to produce segmentation on 3D cloud points, which is different from our project that produce segmentation on 2D RGBD images. The two task is nevertheless similar, and this dataset could potentially be used to evaluate our method. However, it is hard to render realistic images from 3D cloud points directly. Transferring part segmentation from 3D cloud points to 3D models, which consist of surfaces, requires a lot of engineering. In the future, we could complete this transferring process and evaluate our method quantitatively.

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[2] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[4] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[5] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.

[6] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[7] Amir R Zamir, Tilman Wekel, Pulkit Agrawal, Colin Wei, Jitendra Malik, and Silvio Savarese. Generic 3d representation via pose estimation and matching. In *European Conference on Computer Vision*, pages 535–553. Springer, 2016.

[8] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[9] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[10] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.