Pose-Guided Knowledge Transfer for Object Part Segmentation

Shujon Naha Qingyang Xiao Prianka Banik Md Alimoor Reza David J. Crandall
Luddy School of Informatics, Computing, and Engineering
Indiana University
{snaha,mdreza}@iu.edu, {xiaoq,djcran}@indiana.edu, prianka.banik.buet@gmail.com

Abstract

Object part segmentation is an important problem for many applications, but generating the annotations to train a part segmentation model is typically quite labor-intensive. Recently, Fang et al. [6] augmented object part segmentation datasets by using keypoint locations as weak supervision to transfer a source object instance's part annotations to an unlabeled target object. We show that while their approach works well when the source and target objects have clearly visible keypoints, it often fails for severely articulated poses. Also, their model does not generalize well across multiple object classes, even if they are very similar. In this paper, we propose and evaluate a new model for transferring part segmentations using keypoints, even for complex object poses and across different object classes.

1. Introduction

While much work has studied segmenting objects from image backgrounds, the more challenging problem of fine-grained object *part segmentation* would benefit many applications from fine-grained object classification [22], to pose estimation [3, 18], to object re-identification [4], etc. The goal of part segmentation is to produce pixel-level semantic annotations that indicate individual object parts.

Recent work using deep learning has shown impressive performance on object part segmentation for both rigid and non-rigid objects [7, 8, 12, 15]. Most of these papers require large quantities of annotated training images with fine-grained, pixel-wise part segmentation masks, which can be extremely labor-intensive to produce.

Recently, Fang et al. [6] showed that it is possible to generate pixel-level part annotations for an unlabeled target object instance by using keypoints to propagate part segmentations from a labeled source object instance of the same class. This significantly accelerates creating pixel-wise part segmentation masks, since manually annotating keypoint locations is significantly less labor-intensive. While the idea is promising, their work requires that the source and target objects have clearly visible keypoints and very similar poses. Such constraints restrict the usage of their model for many scenarios, e.g., when there are very few annotated source objects and the target objects have very different poses from the source objects.

We thus propose a new model which first directly generates a pseudo-part segmentation only from the object keypoints, and then later combines it with appearance information for improved object part segmentation. In contrast to [6], which requires the source and target instances to have the same number of visible keypoints, our approach can use instances with varying numbers of visible keypoints. Also, as our model directly learns a pose-to-part generation model, it can better generalize to novel poses in the target dataset. Moreover, we also use the fact that many object classes share similar semantic parts, even if their overall appearances are quite different, and thus can be used to augment the annotated dataset for improved performance. For example, while different quadruped (four-legged) animals have widely different sizes and appearances, many share similar body parts and body structures. Thus we can augment the annotated dataset of individual quadruped animal part segmentations by considering different quadruped animals, such as dogs, cats, horses, sheep, etc., as a single class (i.e. quadruped) and improve the part segmentation performance for all of them.

In summary, we propose a new approach for transferring object part annotations from source objects to target objects using keypoint guidance. Through extensive experiments, we show that our approach can handle large variations in the source and target objects, and produce better-quality part segmentation results than existing approaches.

2. Related Work

Part Segmentation With Pose. The strong spatial relationship between object keypoints and parts is used in several papers to improve the accuracy of both pose and part predictions. Xia et al. [20] combined intermediate semantic part score maps with pose estimates to refine part segmentation results. Nie et al. [13] proposed a mutual feature-sharing mechanism between two separate pose and part pre-

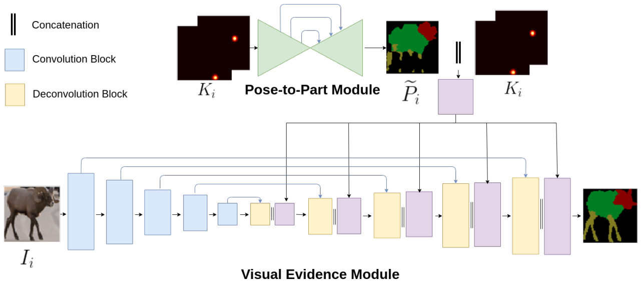


Figure 1: Pipeline of our approach. The Pose-to-Part module first takes the keypoint heatmaps K_i of the current object as input to generate estimated initial part segmentations $\widetilde{P}_i$. Then $\widetilde{P}_i$ and K_i are concatenated and passed to the Visual Evidence module, which uses both the input image I_i and structural information in the initial part segmentations to produce the final segmentation result.

diction networks to improve each others' accuracy. The most relevant work to ours is that of Fang et al. [6] showing that part segmentation of one object can be transferred to another object of the same class using pose guidance.

Weakly Supervised Semantic Segmentation. Using weak supervision for semantic segmentation is a highly studied topic. Class labels [1, 5, 23], point supervision [2], scribbles [11], and bounding boxes [9, 10] are among the most common weak supervisory cues for semantic segmentation. Recently Yang et al. [21] proposed an iterative refinement approach to transform pose-based part priors to full human body part segmentations. However, there is little work that explores using keypoints for generalized multi-class object part segmentation.

3. Our Approach

Our task is to segment parts of an object instance, given only the image and 2D keypoint locations of that instance at inference time. During training, we are given images with both keypoint and part annotations. Assume all training and test objects share a maximum of p body parts and k keypoints. Denote a training instance as $s_i = \{I_i, K_i, P_i\}$, $i = 1...N$, where $I_i \in R^{h\times w\times 3}$ is an input image, $K_i \in R^{h\times w\times k}$ is the heatmap generated from the set of k 2D keypoint annotations, and $P_i \in R^{h\times w\times p}$ is the corresponding pixel-level part segmentation map. Let N be the total number of training images. Consider a test instance as $x = \{I, K\}$. Our goal is to use the provided keypoint annotations to transfer part segmentation labels from the fully annotated training set to a weakly labeled test set.

Our model consists of two main parts: the Pose-to-Part module and the Visual Evidence module. The Pose-to-Part module learns to convert the keypoints K_i of a training instance s_i to an estimated pseudo part segmentation as similar as possible to the actual part annotation P_i of that instance. The Visual Evidence module takes the target image as input and combines the image features with the pseudo part segmentation and the target keypoints K_i to generate the final part segmentation result. An overview of the complete approach is in Figure 1.

3.1. Pose-to-Part Module

The goal of this module is to estimate part segmentations of unseen objects using keypoint annotations only. Keypoints provide useful structural information that can be directly used to estimate part segmentation. Considering that a test object can have quite different shape and number of visible keypoints than the training objects, learning such a model can help to produce generalized part annotations of test objects using their poses. This avoids the constraint of [6] that requires strictly similar poses for the source and target objects.

Pose-to-Part is a U-Net-like [14] network consisting of a fully convolutional encoder-decoder network with skip connections. The encoder reduces the spatial dimensions of the input so that the network can understand the relative locations of the keypoints and the decoder then generates a higher-resolution version of the part annotations.

3.2. Visual Evidence Module

While the Pose-to-Part module can estimate part annotations for the target object, these estimates may be inaccurate due to occlusion, different sizes of the parts, very sparse keypoints, etc. Thus we incorporate visual evidence in addition to the structural evidence from the Pose-to-Part module, by introducing a Visual Evidence network which takes three inputs: (1) the input Image, I_i, (2) the keypoint heatmap, K_i, and (3) the pseudo part annotation output, $\widetilde{P}_i$, from the Pose-to-Part module.

The Visual Evidence network is also a fully-convolutional encoder-decoder network with skip connections. The network first encodes I_i as a convolutional feature map and then passes it through a series of learnable deconvolution layers to predict the final part segmentation output. We also generate multi-scale channel-wise concatenated maps K_i and $\widetilde{P}_i$, and then concatenate them with the visual features at each stage of the decoder module. This allows both the visual and structural features to refine each other and produce a better joint part segmentation result than either could do alone.

3.3. Training

The network is trained end-to-end using two loss functions. The first loss is calculated between the final output of the Pose-to-Part network $\widetilde{P}_i$ and the ground truth part segmentation,

$$L_{trans} = \sum_i \sum_j e[\widetilde{P_i(j)}, P_i(j)], \tag{1}$$

where j is the pixel index. $P_i(j)$ is the ground truth part annotation for target i, and e indicates the per-pixel cross entropy loss.

The second loss L_{seg} is used at the end of the Visual Evidence module to calculate the loss for the final part segmentation result,

$$L_{seg} = \sum_i \sum_j e[V^j(I_i, K_i, \widetilde{P}_i; \theta), P_i(j)], \tag{2}$$

where V is the Visual Evidence module, and I_i and K_i are the image and keypoint heatmaps of the i-th target. The final loss combines both of these,

$$L = \lambda_{trans} * L_{trans} + \lambda_{seg} * L_{seg}, \tag{3}$$

where λ_{trans} and λ_{seg} are weights.

4. Experiments

We present findings from extensive experiments to evaluate the effectiveness of the proposed method.

4.1. Dataset

Pascal Part [19] is a part segmentation dataset with pixel-wise object part annotations, keypoint locations, and bounding box annotations. The dataset includes five quadruped animals, `Cat`, `Cow`, `Dog`, `Horse`, and `Sheep`. We use the subset of images having at least one of these objects. We use the bounding box labels to crop the objects, discarding bounding boxes where there is overlap with another bounding box with an IoU of more than 0.05. We also discard bounding boxes that have any side smaller than 32 pixels or for which the object has less than 5 keypoints. After applying this filter, we get in total 2872 images of quadrupeds from these five classes (245 `Sheep`, 404 `Horse`, 233 `Cow`, 1097 `Dog`, 893 `Cat` images). While the dataset contains more detailed part annotations, we follow the previous work [6] and only consider four parts for each animal: *head, torso, legs,* and *tail*.

4.2. Implementation Details

For the Pose-to-Part network, we first convert the keypoint annotations into heatmaps using a Gaussian function with $\sigma = 7$. The encoder of the Pose-to-Part network has 5 downsampling residual blocks and the decoder has 5 upsampling residual blocks. The upsampling blocks use PixelShuffle layers [16] for the upsample operations. For the Visual Evidence module, we use the encoder-decoder network with skip-connections from [17]. The encoder consists of an ImageNet-pretrained VGG-16 network, and the decoder consists of a series of 5 upsampling blocks with learnable deconvolution layers. All the layers in the image evidence are learned during training. We train the full network end-to-end and use $\lambda_{trans} = 0.01$ and $\lambda_{seg} = 1$. We use batch size 24 and resize the input images and the ground truth part segmentations to 256×256 during training.

4.3. Baselines

RefineNet is the affine transformation-based approach proposed by Fang et al. [6]. It requires nearest neighbors based on pose similarity to perform the morphing for body part parsing. We follow their settings to train the model.

Transform is our Pose-to-Part module. For this baseline, we disable the Visual Evidence module and consider $\widetilde{P_{t^{c'}}}$ as our final output.

TernausNet [17] is the basic encoder-decoder network used as our Visual Evidence module. We only consider the image as input without keypoint locations for this baseline.

4.4. Evaluation on Pascal Part Dataset

We first train on one animal class and test on the same animal class, as in [6]. We randomly choose 80% of the images for training and 20% for testing on each class. All experiments use 5-fold cross-validation. Figure 2 presents the results using intersection-over-union (IoU) as the evaluation metric. The figure shows that across all animal classes, our model outperforms most baselines on particular body parts, and outperforms all baselines averaged across all body parts. This confirms that the Pose-to-Part module is indeed adding useful information to the network.

We find that RefineNet [6] has the worst performance among all the baselines. RefineNet can achieve 39.80%, 36.85%, 32.38%, 31.09%, 28.23%, 17.27% on `Sheep`, `Horse`, `Cow`, `Dog`, and `Cat` respectively, in terms of IoU averaged across all parts. In contrast, our full model achieves 49.83%, 60.16%, 51.48%, 59.50%, and 58.84% on `Sheep`, `Horse`, `Cow`, `Dog`, and `Cat`. From qualitative results we observe that RefineNet often struggles to predict accurate pixel labels (Figure 3a). Since it heavily relies on source and target objects having similar poses, it fails to make accurate predictions if the same keypoints are not visible in the source and target objects. These results suggest that our model performs better because it can better utilize keypoint annotations than RefineNet.

Transform, our second baseline, performs slightly better than RefineNet but still fails to segment the small parts (such as *tail*). It always performs much better than RefineNet for *head*, probably because the head has more keypoint annotations than the other parts. Dense keypoints result in better pseudo-part annotation generation from the Pose-to-Part module. This indicates that more keypoint annotations can help to improve the performance of this module. Our full model performs much better than the Transform baseline, indicating that both the Pose-to-Part module and Visual Evidence modules play crucial roles in producing high-quality part segmentation results.

TernausNet [17], our third baseline, performs much better than RefineNet and Transform, suggesting that appearance is highly important for recognizing object parts. Although our model sometimes performs similarly or only

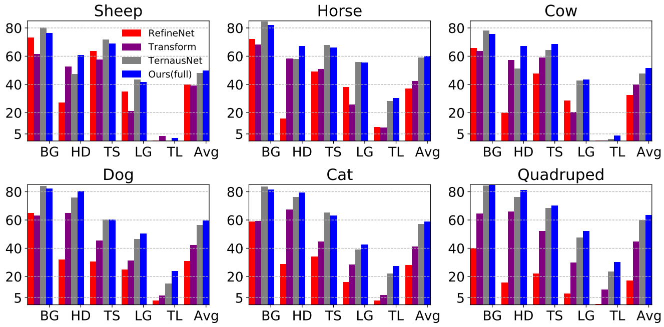


Figure 2: Evaluation on Pascal Part dataset in terms of 4-part parsing. For the first five plots (`Sheep` through `Cat`), we train on one animal class and test on the same animal. For the last case (`Quadruped`), we combine all the images from the five classes and consider them as a single class (`Quadruped`) for both training and testing. Parts are BG for *background*, HD for *head*, TS for *torso*, LG for *legs*, TL for *tail*, and Avg is average of all. Best viewed in color.

slightly better than TernausNet in terms of average IoU, it always produces better results for the smallest part (i.e., *tail*), especially for animals like `Dog` and `Cat`. These parts and classes have the most articulated poses and thus are among the most challenging.

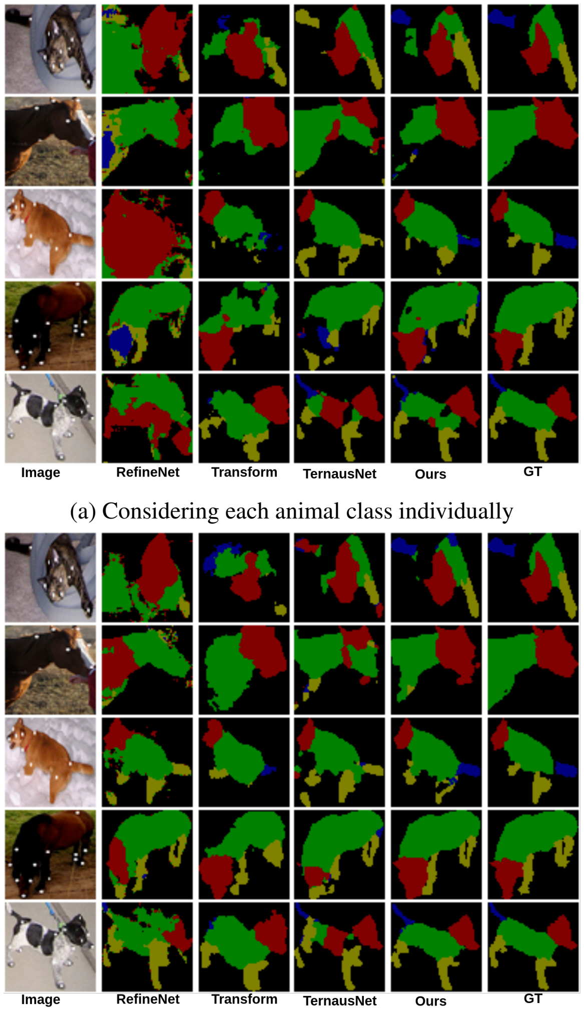
Interestingly, all of these baseline models perform significantly better on `Horses`, presumably because `Horse` has less pose diversity than the other classes.

An advantage of our approach is that we can transfer part segmentations *across* object classes, which could lead to better results by effectively augmenting the annotated dataset with more diversity in terms of animal sizes, shapes, and poses. We evaluate this by considering all quadruped animal classes as a single class. As shown in Figure 2, our model achieves 63.69%, which is significantly higher than our results on any of the individual classes (which ranged from 49.83% to 60.16%). This indicates that our model can utilize the part annotations from multiple quadruped animal classes. In contrast, the performance for RefineNet dropped significantly, to 17.27%. Our results on `Quadruped` also significantly outperform those of TernausNet (60.00%), suggesting that our model is better at generalization.

Qualitative results of our models and the baselines are presented in Figures 3a and 3b. We see that after considering all animals as `Quadruped`, our model's performance improves significantly compared to the baseline models.

5. Conclusion

In this paper, we explore the problem of object part segmentation with limited data. We propose a novel approach to transfer part annotation from a labeled set to an unlabeled set using keypoint annotations as transfer guidance. We show that our approach can help to produce better part segmentation for both the single object class and joint object class settings. We hope our work will lead to more research on the problem of generalized part segmentation across multiple classes.



(a) Considering each animal class individually

(b) Considering all animals as a single class, `Quadruped`

Figure 3: Qualitative comparison on Pascal Part dataset.

Acknowledgments. This work was partially supported by the National Science Foundation (CAREER IIS-1253549) and the Indiana University Office of the Vice Provost for Research, the College of Arts and Sciences, and the Luddy School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project *Learning: Brains, Machines, and Children.*

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised

semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4981–4990, 2018.

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 549–565. Springer, 2016.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

[4] 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 (CVPR)*, pages 1335–1344, 2016.

[5] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 642–651, 2017.

[6] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018.

[8] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 932–940, 2017.

[9] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4233–4241, 2018.

[10] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 876–885, 2017.

[11] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016.

[12] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 909–918, 2019.

[13] Xuecheng Nie, Jiashi Feng, and Shuicheng Yan. Mutual learning to adapt for joint human parsing and pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 502–517, 2018.

[14] 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.

[15] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4814–4821, 2019.

[16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[17] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018.

[18] Min Sun and Silvio Savarese. Articulated part-based model for joint object detection and pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 723–730. IEEE, 2011.

[19] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. Joint object and part segmentation using deep learned potentials. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1573–1581, 2015.

[20] Fangting Xia, Peng Wang, Xianjie Chen, and Alan L Yuille. Joint multi-person pose estimation and semantic part segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6769–6778, 2017.

[21] Zhengyuan Yang, Yuncheng Li, Linjie Yang, Ning Zhang, and Jiebo Luo. Weakly supervised body part parsing with pose based part priors. *arXiv preprint arXiv:1907.13051*, 2019.

[22] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 834–849. Springer, 2014.

[23] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3791–3800, 2018.