TransFGU: A Top-down Approach to Fine-Grained Unsupervised Semantic Segmentation

Zhaoyuan Yin[1,2*], Pichao Wang[2†], Fan Wang[2], Xianzhe Xu[2], Hanling Zhang[3‡], Hao Li[2], and Rong Jin[2]

[1] College of Computer Science and Electronic Engineering, Hunan University, China
zyyin@hnu.edu.cn
[2] Alibaba Group, China
{pichao.wang, fan.w, xianzhe.xxz, lihao.lh, jinrong.jr}@alibaba-inc.com
[3] School of Design, Hunan University, China
jt_hlzhang@hnu.edu.cn

In this supplementary material, we first present more training details. Then we present more ablation studies to reveal the effectiveness of K, the semantic prior model used in our pipeline, the cropping manner and the foreground prior. Our re-defined labels for 5 and 16 categories in LIP are also presented. Finally, we show failure cases, the visualization of CAM and more qualitative results on Cityscapes and LIP.

Training Details. We train our model on 4 × Tesla V100 32G GPUs with Adam optimizer. In general, a larger batch-size benefits the learning stability, and a larger amount of images and target classes requires a larger learning rate and training epochs. Different learning rates, batch-size, and epochs are used for the four benchmarks for a better trade-off between the performance and the GPU memory cost. Specifically, we set 256/1e-4/200 (batch size/learning rate/epochs) for MS-COCO, 512/1e-4/100 for PascalVOC, 256/5e-5/50 for Cityscapes and 256/2e-5/100 for Lip. The learning rate is fixed during the whole training process. The α in peer loss linearly increases from 0.03 to 0.1 according to the number of training epochs. The background threshold T^{bg}, loss weights $\omega 1, \omega 2$, and the parameter β in the step size of sliding window is fixed across all the benchmarks.

The Effectiveness of K. We perform overclustering on the Pascal-VOC to reveal the effectiveness of different settings of K to a certain target semantic granularity level. Specifically, we set different K which larger than the number of ground-truth categories as the target amount of cluster when performing K-Means to the top-level semantic features, and use the Hungarian algorithm to find the greatest matching. Note that the unmatched categories are discarded when calculating the mIoU and accuracy. The results shown in Table 1 indicate a reasonable K value that closed to the amount of ground-truth categories can benefit the evaluation performance, while an over-large K leads to the performance drop. It is due to the over-clustering in top-level semantic features leads to

* Work done during an internship at Alibaba Group.
† Corresponding author, project lead.
‡ Corresponding author.

a finer semantic division that would divide a whole object into its sub-categories, and causes the mismatching between predicted and the ground-truth semantic clusters. We emphasize that trying to find an 'objective' value of K concerning a certain validation set in an unsupervised way is difficult. It's due to the intrinsic non-uniqueness and hierarchy of semantic definition, e.g. both 'left/right arm' or 'arms' are acceptable semantic divisions on the same validation set, depending on which granularity level is required. To this end, the introduction of K in our pipeline is a necessary and reasonable way to enable flexible control on any desired granularity levels.

Table 1. Over-clustering Results on Pascal-VOC (20 ground-truth categories) with different setting of K.

K	20	22	25	30	50
mIoU	37.15	37.26	37.68	29.39	27.42
Acc	83.59	84.12	82.65	80.85	75.32

Table 2. Results for the effectiveness of semantic prior model.

SSL method	patch size	mIoU	Acc.
DINO	8	11.93	34.32
DINO	16	9.43	28.12
MoCoV3	16	2.32	20.94

The Effectiveness of the Semantic Prior Model. We conduct two more experiments to reveal the effectiveness of the semantic prior model. The results are shown in Table 2. We first change the semantic prior model from ViT small 8x8 to ViT small 16x16 trained by DINO [1]. The mIoU drop from 11.93 to 9.43. Second, we change the ViT small 16x16 trained by DINO with the one trained by MoCo V3 [4]. The performance drops dramatically to 2.32, which indicates the different manners of self-supervised learning affect the segmentation property of the semantic prior model and result in different performances in our pipeline.

The Quality of Cropping. An extra experiment is conducted by using ground-truth labels instead of sliding windows to crop the foreground and background objects in a more precise way, which can be regarded as an upper limit of our cropping operation. The mIoU on COCO-Stuff-171 is slightly improved from 11.93 to 12.16. This shows that, even though the sliding window cropping cannot generate precise boxes around semantic objects, it still has exploited sufficient semantic information, leaving not much room to improve compared with the ground-truth cropping.

The Foreground Prior. Experiments of clustering the semantic features without separating the patches with foreground prior will result in the drop of mIoU from 11.93 to 9.52 on COCO-Stuff-171, indicating that the separation of foreground and background patches is helpful to improve the quality of semantic feature clustering.

Label Definition on LIP under Different Granularity. We re-defined the LIP [3] under different granularity by merging categories in its original label (19 categories) and generating two new label definitions (16 and 5 categories). The indices projection is shown in Table 3.

Table 3. Label definition on LIP under different granularity.

Name	granularity level		
	19 categories	16 categories	5 categories
Background	0	0	0
Hat	1	1	1
Hair	2	2	1
Glove	3	3	3
Sunglasses	4	4	1
Upper-clothes	5	5	2
Dress	6	6	2
Coat	7	7	2
Socks	8	8	5
Pants	9	9	4
Jumpsuits	10	10	2
Scarf	11	11	2
Skirt	12	12	2
Face	13	13	1
Left-arm	14	14	3
Right-arm	15	14	3
Left-leg	16	15	4
Right-leg	17	15	4
Left-shoe	18	16	5
Right-shoe	19	16	5

Failure Cases. Since our segmentation is based on the high-level semantic representation, similar categories may confuse the classification (The two cases in the left column of Figure 1). Besides that, other challenging cases would also cause failure of our method, *e.g.* occlusions and tiny objects (top-right and bottom-right in Figure 1).

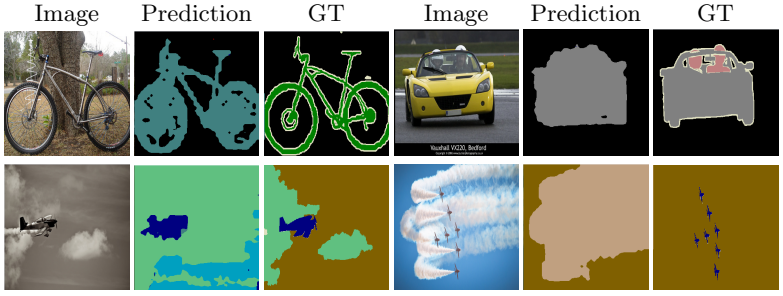

Fig. 1. Failure Cases. left: misclassified; right: missegmented.

CAM Visualization. Figure 2 visualizes the class activate map generated by Grad-CAM. As one can see, the CAM can locate objects well in a very complex scene-based image and give a fine-grained activate map of the target location, which is crucial to our top-down segmentation pipeline.

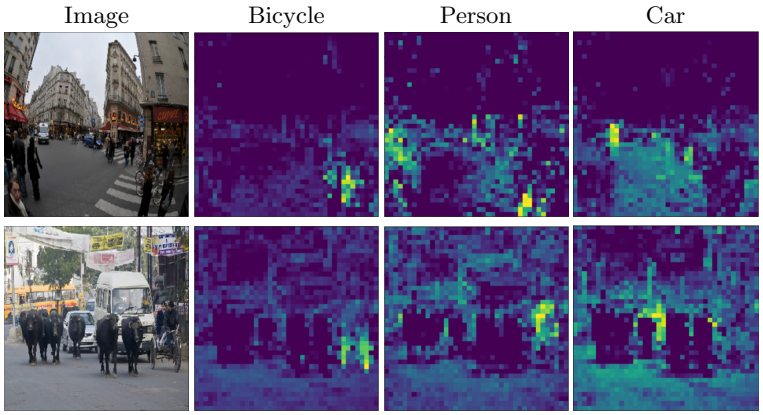

Fig. 2. CAM visualization. We show three CAMs correspondence to three particular semantic classes, *i.e.* bicycle, person and car in the same image.

Qualitative Results. We show more qualitative results on Cityscapes [2] and LIP [3] in Figure 3 and Figure 4, respectively.

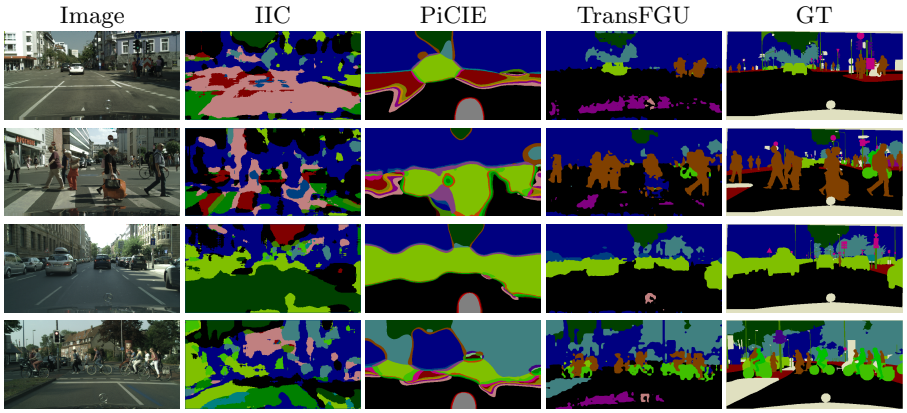

Fig. 3. Qualitative comparison on Cityscapes.

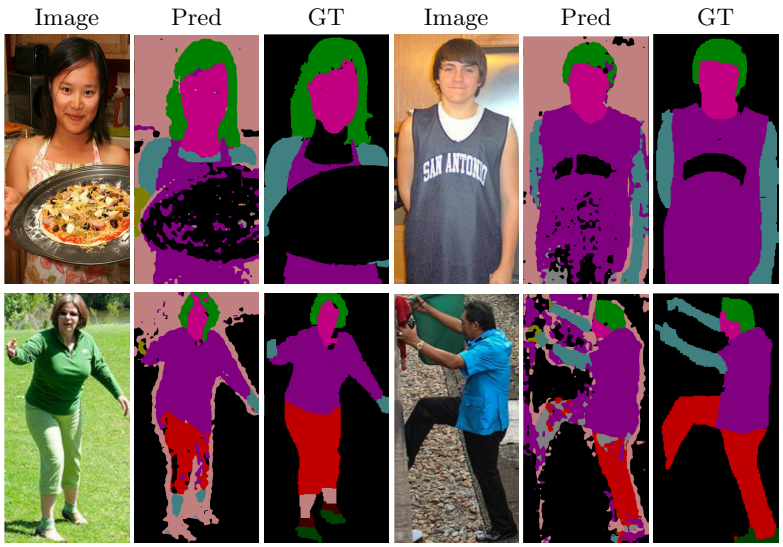

Fig. 4. Qualitative results on LIP for 16 fine-grained semantic granularity.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
3. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: 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. pp. 932–940 (2017)
4. Yao, Z., Cao, Y., Lin, Y., Liu, Z., Zhang, Z., Hu, H.: Leveraging batch normalization for vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 413–422 (2021)