MatchMask: Mask-Centric Generative Data Augmentation for Label-Scarce Semantic Segmentation

Supplementary Material

A. More Implementation Details

Semantic image synthesis: During training, we resize all semantic masks to 512x512 following FreestyleNet [11]. We freeze the entire model and only fine-tune the proposed Adapter with LoRA rank set to 4. The training can be completed within 27 hours for 100k iterations on a single A100 GPU. In the mask-to-image synthesis stage, the output images are resized to the original resolution of the input masks, with each mask generating 5 images as default.

Semantic segmentation model training. All the semantic segmentation models are trained in the mmsegmentation codebase. For VOC and COCO, we adopt DeepLabv3+ [2] model with ResNe101 [5], and Mask2former [3] with Swin-B [7] for ADE20K dataset. Images are randomly scaled to [0.5, 2.0] and cropped to 512x512. The batch size is set to 8 for each GPU, and a total of 2 GPUs are used. We use AdamW optimizer with 1e-4 weight decay. The initial learning rate is 1e-4, with the polynomial learning rate decay $lr_{iter} = lr_{init}(1 - \frac{iter}{maxiter})^{\gamma}$, where $\gamma = 0.9$. In the data-limited setting, we train VOC (92/182/366/732) for 3k/5k/8k/10k iterations, train COCO for 20k iterations and train ADE for 10k iterations. For the semi-supervised setting, we extend training iterations to 20k for VOC and ADE. We use default image augmentation in mmsegmentation, such as RandomCrop, RandomFlip and PhotoMetricDistortion.

Methods	baseline	Cutout	Mosaic	MatchMask
mIoU	67.5	68.1	67.6	**72.1**

Table S1. Comparison to advanced image transforms.

B. Additional Experiments

B.1. Comparison with advanced image transforms

In this section, we compare MatchMask with stronger image-level data augmentation, including Cutout [4] and Mosaic [1]. Cutout randomly drops some regions of the image. Mosaic combines the four images given into one output image. The output image is composed of the parts from each sub-image. We compare them to our MatchMask on VOC with 1/4 (366) labeled data. Results in Table S1 indicate that the performance gains from these two methods are minimal and significantly inferior to our solution, validating the superior ability of MatchMask in the label-scarce scenario.

Methods	Params.	FID ↓	DINO-SIM ↑
Few Samples (1%):			
UNet	859M	29.1	44.7
Residual block	615M	32.2	40.1
Transformer block	242M	27.9	45.6
Critical layer	72M	**27.4**	46.2
Ours	0.7M	27.6	**48.0**
All Samples:			
UNet	859M	**26.2**	49.4
Ours	0.7M	27.1	**49.4**

Table S2. Quantitative results on ADE20K in few-sample setting.

B.2. Comparison with other fine-tuning strategies

Baseline. Since our Adapter is the extension of FreestyleNet in the few-shot setting, we treat it as the baseline, which fine-tunes the entire U-Net. In addition, we incorporate three common fine-tuning methods: (1) fine-tuning residual blocks in U-Net; (2) fine-tuning transformer blocks in U-Net; (3) fine-tuning critical layers selected by our Gradient Probe Method. To be consistent with FreestyleNet, we conduct experiments on ADE20K, randomly selecting 200 samples (about 1%) as labeled data and evaluating on the validation set to verify whether the model is overfitting.

Evaluation Metrics. To ensure comparability with FreestyleNet, we adopt Fréchet Inception Distance (FID) [6] to assess the visual quality of generated images. To measure layout consistency, unlike previous methods using a pre-trained segmentation model (*e.g.*, UperNet101 [10]) on the specific dataset to calculate mIoU between GT and predictions, which suffers from model's low performance and limited generalization, we propose a more effective evaluation metric termed DINO-SIM. It is based on DINOv2 [9], which is a large pre-trained self-supervised model and has powerful representation capabilities. The pixel-level features learned by DINOv2 are well-suited for dense prediction tasks such as semantic segmentation. We use DINOv2 to extract dense features from the generated/original images (resized to high resolution 840×840) and compute the cosine similarity between these features to measure layout consistency. This method provides improved accuracy and generalization compared to traditional metrics.

Results. In Table S2, we can observe that 1) Transformer blocks play a more significant role in spatial information control than residual blocks, even though they have fewer trainable parameters; 2) The identified critical layers prove

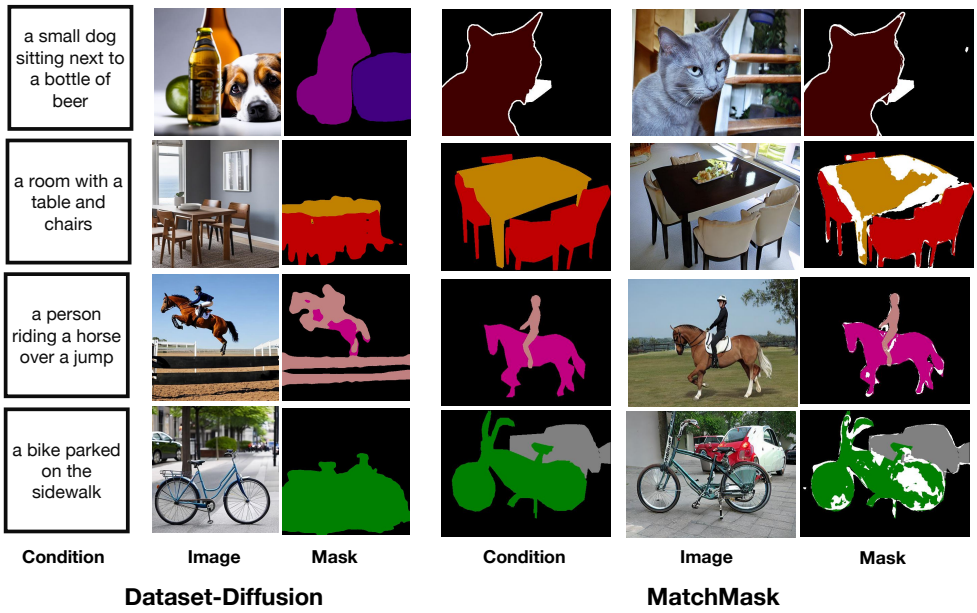

Figure S1. Visualizations of text-centric and mask-centric generative data augmentation.

to be effective, yielding better results than incorporating task-irrelevant layers; 3) Our Adapter attains better layout consistency (higher DINO-SIM) with a remarkably small number of parameters; 4) When using all samples, our model's performance continues to improve, demonstrating a scalable capability. With only 1% of the training data and 0.7 million trainable parameters, our method achieves performance comparable to fully-supervised methods.

C. Discussions and Limitations

Discussions: In this work, we propose a novel mask-centric generative data augmentation paradigm for label-scarce semantic segmentation. We have validated its superior performance on various benchmarks. The key design is the Adapter to achieve mask-to-image synthesis via only few densely labeled samples, which is based on the pre-trained Stable Diffusion model. We believe the semantic image synthesis process can be further enhanced by adopting more powerful generative models in the future. In addition, we directly produce K images for each mask in our experiments while not focusing on the value of each sample for the model training. A potential improvement direction is to select more hard samples for segmentation models (*e.g.*, the classes showcase inferior performance).

Limitations: While MatchMask has shown impressive performance for data augmentation, it still has some limitations. First, the inference speed during image synthesis is unsatisfactory (*e.g.*, 4s/image on VOC or 6s/image on ADE) using PLMS sampling for 50 steps. This stems from the inherent drawbacks of the diffusion model, and we believe future advances in generative paradigms or samplers could address this. Additionally, although MatchMask could produce images based on the pseudo mask, the synthetic image would become messy and unrecognizable if the mask is too noisy, resulting in confusion for model training. The potential solutions include combining with semi-supervised works to obtain more accurate pseudo masks or designing some principles to neglect these unreliable samples.

D. More Visualizations

In Fig S1, we provide additional qualitative results generated by our mask-centric MatchMask and text-centric Dataset-Diffusion [8]. We can observe that 1) Dataset-Diffusion sometimes generates delusive images that are not in the same domain as labeled images. This is the intrinsic limitation of text-centric techniques because it is not trivial to ensure it via text alone. In contrast, MatchMask fine-tunes the proposed Adapter on a few source images, yielding in-domain images. 2) For the synthetic image-text pairs, MatchMask showcases higher quality than Dataset-Diffusion. This is because Dataset-Diffusion follows a *generate image then annotate it* manner, which leverages cross-attention between the words in the text and regions in the generated image to produce mask annotation, yielding inferior performance. Differently, MatchMask follows the *generate image to match given mask* manner, leading to more reliable training data.

In Fig S2, we present the multiple images generated by the same mask on VOC, COCO and ADE. Results show that MatchMask can produce images with diverse contents and styles, which is beneficial for data augmentation.

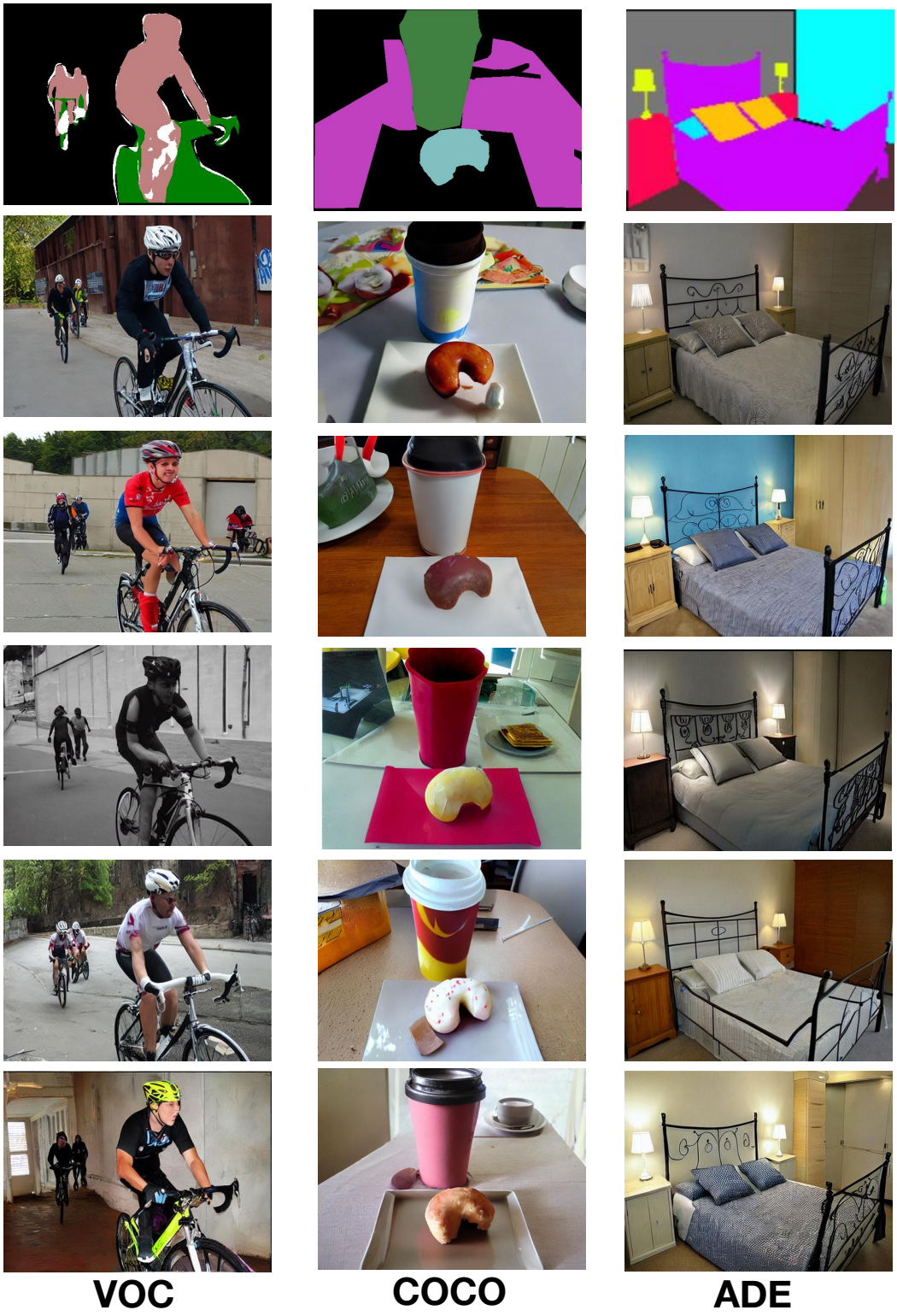

Figure S2. Visualizations of diverse synthetic images on VOC, COCO and ADE.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 1

[4] Terrance DeVries. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[8] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[9] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[10] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 1

[11] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023. 1