FreeMask: Synthetic Images with Dense Annotations Make Stronger Segmentation Models

Lihe Yang[1] **Xiaogang Xu**[2,3] **Bingyi Kang**[4] **Yinghuan Shi**[5] **Hengshuang Zhao**[1*]
[1]The University of Hong Kong [2]Zhejiang Lab [3]Zhejiang University
[4]ByteDance [5]Nanjing University
https://github.com/LiheYoung/FreeMask

A More Implementation Details

Following FreestyleNet [1], we resize all semantic masks to 512x512 for image synthesis. The guidance scale of the diffusion model is set as 2.0, and the sampling step is 50. For better reproducibility, we pre-define and fix a sequence of $N_{\max}$ seeds to synthesize densely annotated images.

For synthetic pre-training, we adopt exactly the same training protocols as real images. Then, during fine-tuning, the base learning rate is decayed to be half of the normal learning rate. Since our whole model parameters are pre-trained with synthetic images, the fine-tuning learning rate is the same throughout the whole model. The model is pre-trained and fine-tuned for the same iterations as real images. For joint training with real and synthetic images, real images are over-sampled to the same scale as synthetic images to make each mini-batch evenly composed of real images and synthetic images. The batch size is the same as real-image training. Each real image is iterated for the same number of epochs as the real-image training paradigm.

As for other hyper-parameters, *e.g.*, data augmentations and evaluation protocols, they are set exactly the same as those in regular training paradigms. We adopt the MMSegmentation codebase for our development. We use 8 $\times$ Nvidia Tesla V100 GPUs for our training experiments.

B The Most Improved Classes

We list the most improved ten classes on ADE20K (the gain is measured by IoU): (1) ship: +68.19, (2) microwave: +48.72, (3) arcade machine: +45.85, (4) booth: +45.66, (5) oven: +30.86, (6) skyscraper: +23.23, (7) swimming pool: +15.52, (8) armchair: +14.6, (9) hood: +14.43, (10) wardrobe: +13.24.

C Discussions for Future Works and Limitations

Future works. In this work, we use the off-the-shelf semantic image synthesis model to generate densely annotated images. We have validated that the fully-supervised baseline can be remarkably boosted with these synthetic training pairs. We expect more considerable improvements can be achieved in future works by (1) better-trained or larger-scale pre-trained generative models, (2) larger-scale synthetic training pairs, and (3) taking the class distribution into consideration during image synthesis.

Limitations. It is relatively time-consuming to produce synthetic training pairs. For example, it takes around 5.8 seconds to synthesize a single image with a V100 GPU. In practice, we speed up the synthesis process with 24 V100 GPUs. Therefore we can construct the entire synthetic training set for ADE20K and COCO-Stuff in two days. In addition, considering the great potential of our densely annotated synthetic images, it will be more practical to apply our proposed roadmap to real-world

*Corresponding author

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

scenarios, *e.g.*, medical image analysis and remote sensing interpretation. We plan to conduct these explorations in future works.

D Visualization of Densely Annotated Synthetic Images and Filtered Regions

We display our diverse densely annotated synthetic images in Figure 1 of ADE20K and Figure 2 of COCO-Stuff. Besides, we also visualize our filtered regions during training for each synthetic image. It can be observed that there exist several patterns of filtered regions, *e.g.*, boundary regions, synthesis failure cases, and small or rare objects. Please refer to the following pages for details.

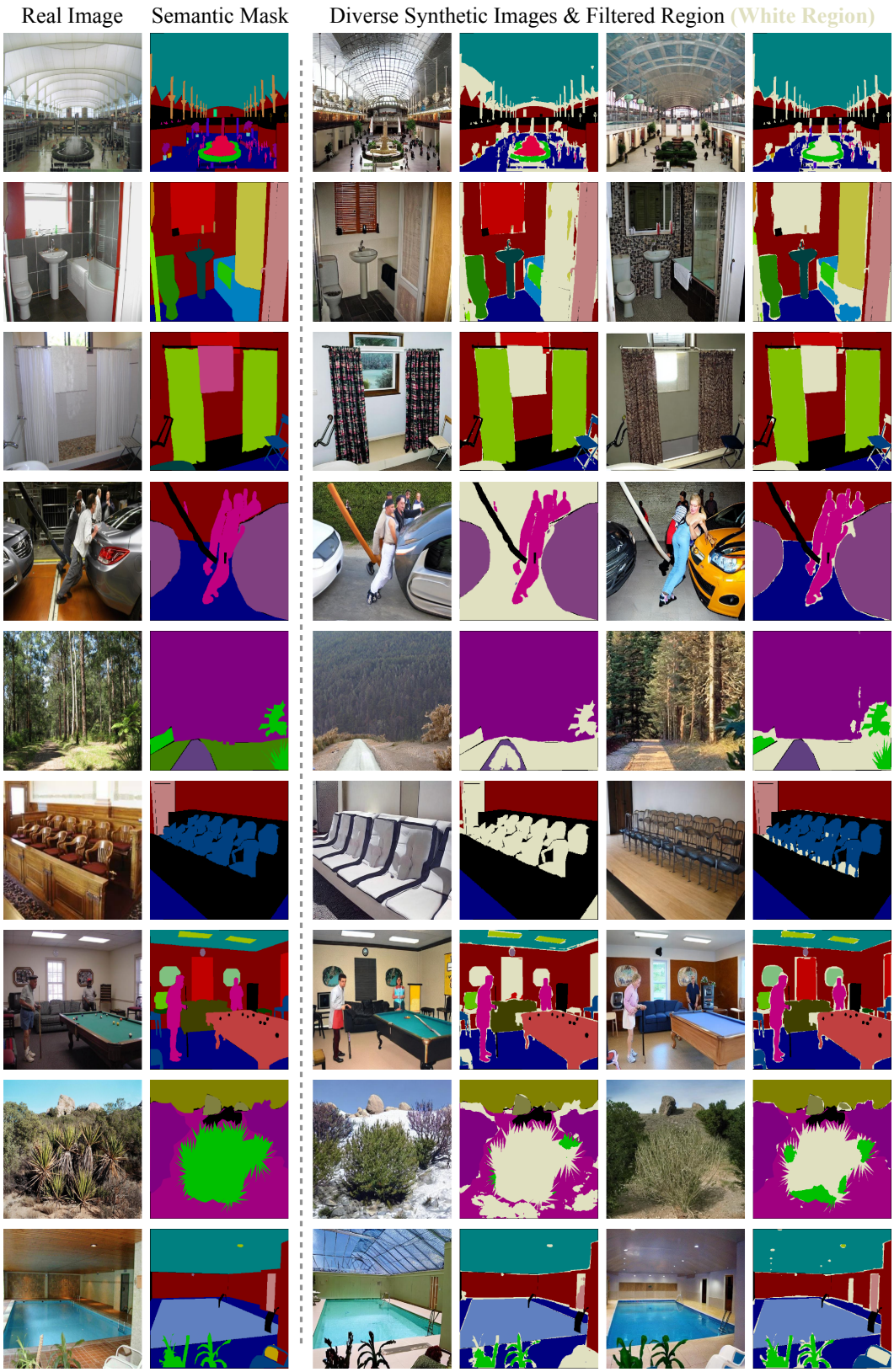

Figure 1: Visualization of diverse densely annotated synthetic images on **ADE20K**, as well as the filtered regions (white regions in the semantic mask). Note that the black regions in the masks are officially marked as "ignored region" by the original ADE20K dataset.

Figure 2: Visualization of diverse densely annotated synthetic images on **COCO-Stuff**, as well as the filtered regions (white regions in the semantic mask).

References

[1] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *CVPR*, 2023. 1