NamedMask: Distilling Segmenters from Complementary Foundation Models Supplementary Material

Gyungin Shin[1,3] Weidi Xie[1,2] Samuel Albanie[3]
[1]Visual Geometry Group, University of Oxford, UK
[2]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, China
[3]Cambridge Applied Machine Learning Lab, University of Cambridge, UK
https://www.robots.ox.ac.uk/~vgg/research/namedmask

In this supplementary material, we describe additional details on our use case of the LAION-5B dataset as an unlabelled image collection (Sec. 1), how we disambiguate homonyms among the ImageNet-S category names (Sec. 2), and prompt engineering considered to extract textual features for image retrieval (Sec. 3). In Sec. 4, we conduct an experiment on the maximum number of images used for copy-paste augmentation and we detail how we decide a background threshold for MaskCLIP and ReCo in Sec. 5. Lastly, we visualise more examples of NamedMask including failure cases (Sec. 6).

1. Preprocessing LAION-5B

As described in Sec. 4.1 in the paper, we use the LAION-5B dataset [3] to build image archives for 919 categories in the ImageNet-S benchmark [2] with CLIP. Preparing a subset of LAION-5B containing the ImageNet-S categories consists of two main steps: downloading images from the official LAION search demo, and filtering human images out from the downloaded images. Each step is detailed in the following.

Downloading images. To download LAION-5B images, we use the official demo page of LAION-5B[1] to download metadata including an image URL. For this, we search each category name with several search options. In detail, to ensure that no inappropriate images are retrieved from the LAION-5B database, we use *Safe mode* and *Remove violence*, which filter out NSFW and harmful images, respectively. We also apply *Hide duplicate urls*, and *Hide (near) duplicate images* functions to diversify retrieved images. Then, we download top 500 images for each category which have the highest similarity score with a given category text embedding by using their URLs in the metadata. We note that there are some URLs which are broken in which cases we simply omit the URL and download the image with the next highest similarity score.

Filtering images with human faces. To avoid using images containing personal information, we utilise a face detector to filter out human images. For this we use ResNet50-based RetinaFace [1] provided by the open source InsightFace package.[2] As a result, we obtain 425,618 images in total, with 463 images for the average number of images per category and 326.0 and 360.5 for an average height and width of the images.

2. Disambiguating ImageNet-S categories

In the main paper, we consider the ImageNet-S benchmark for evaluating our model. We note that there are two pairs of categories sharing a same name with different meanings: *crane* (bird) and *crane* (tower), and *cardigan* (sweater) and *cardigan* (a dog breed). As this can negatively affect the image retrieval step for constructing image archives, we clarify the meaning of the words by changing *crane* (bird), *crane* (tower), *cardigan* (sweater), and *cardigan* (a dog breed) to *crane bird*, *tower crane*, *cardigan sweater*, and *cardigan welsh corgi* respectively, for collecting images from LAION-5B.

3. Prompt engineering

Following [4, 5], we use 85 templates to ensemble textual features for a given category, *e.g.* "`a photo of a category`" and "`a drawing of the category`". In detail, to produce an ensemble of text features for a category, we average and L2-normalise all of the normalised textual features extracted with the 85 templates. The resulting feature is used for retrieving images by comparing to a normalised visual feature from a CLIP image encoder.

[1] https://rom1504.github.io/clip-retrieval

[2] https://github.com/deepinsight/insightface

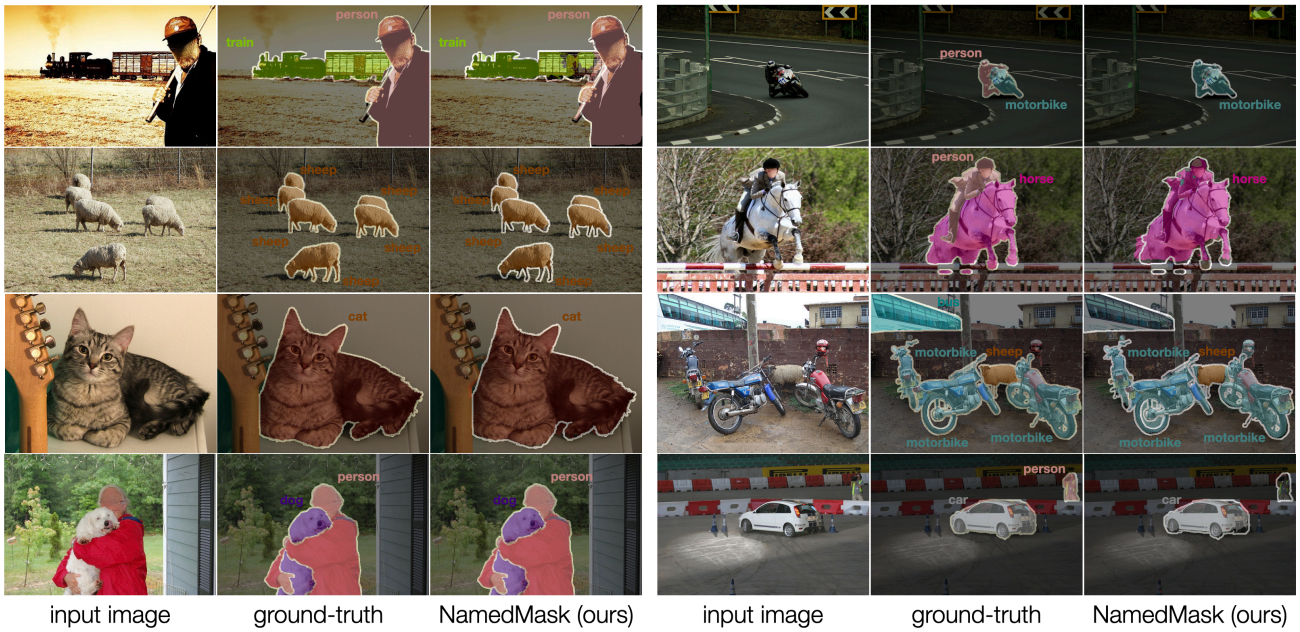

Figure 1. **Qualitative results of NamedMask on VOC2012.** Left: successful cases. Right: typical failure cases. On the ground-truth masks and the predictions, we show a category name of each (predicted) object. Human faces are blurred and ignored regions (*i.e.* object boundaries) are highlighted in white. Best viewed in colour.

copy-paste	n_{max}	mIoU
✗	-	56.6
	2	**58.7**
	4	58.3
✓	6	58.5
	8	58.5
	10	58.5

Table 1. The effect of the maximum number of images used for copy-paste on VOC2012. For reference, NamedMask trained without copy-paste is highlighted in gray.

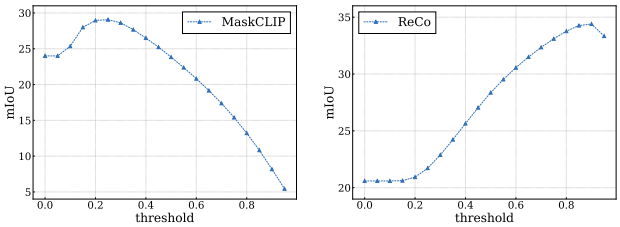

Figure 2. **Finding an optimal background threshold for MaskCLIP (left) and ReCo (right).** For both cases, we vary a threshold from 0.0 to 0.95 with an interval of 0.05 below which is regarded as background. We observe that performance of both methods varies greatly depending on a threshold.

4. Hyperparameter selection for copy-paste

When we consider copy-paste augmentation for training NamedMask, we set a hyperparameter for the maximum number of images n_{max} used for the copy-pate operation. For instance, when n_{max} is set to 10, we randomly select 1 to 10 images to be used for copy-paste at each iteration. It is worth noting that when 1 is selected, it means no copy-paste is applied at the iteration. To investigate the effect of n_{max} on performance of our model, we train NamedMask with different n_{max} values among $\{2, 4, 6, 8, 10\}$ and evaluate on the VOC2012 training split. For reference, we also report a model which is trained without copy-paste.

As can be noted In Tab. 1, while the models trained with copy-paste always show superior performance than the one trained without copy-paste, setting n_{max} to 2 allows for the best performance. For this reason we fix n_{max} to 2 for our models in the main paper.

5. Finding a background threshold

In Sec. 4.4 of the main paper, we compare our model to MaskCLIP [5] and ReCo [4] on VOC2012 and ImageNet-S, both of which contain a background class. As MaskCLIP and ReCo do not explicitly predict a background category, we treat pixels whose maximum class probability is lower than a certain threshold as background. To find the best threshold for each method, we evaluate each model on the VOC2012 training set with a varying threshold t between [0, 0.95] with an interval of 0.05. As shown in Fig. 2, 0.25 and 0.9 allows for the best performance for MaskCLIP and ReCo respectively, we therefore use these thresholds to decide a background pixel throughout our experiments.

6. More visualisation samples

In Fig. 1, we visualise qualitative examples of our model on VOC2012. On the left, we show successful cases whereas on the right, we show typical failure cases.

As discussed in Sec. 5 of the main paper, we note that our model struggles to distinguish a category object and another object of different categories if they tend to co-occur in many cases. For example, a motorbike tends to appear with a person riding it or a horse is inclined to be present with a rider. We conjecture that this is due to our use of a saliency detector for generating (initial) pseudo-masks. That is, pseudo-masks generated by a salient object detector do not separate salient regions based on their semantics, but rather group dominant regions as a whole regardless of their meaning.

To overcome this weakness, it might be helpful to introduce a language-based attention mechanism as considered in [4] for refining pseudo-masks from a saliency model.

References

[1] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 1

[2] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *arXiv:2106.03149*, 2021. 1

[3] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: a new era of open large-scale multi-modal datasets. https://laion.ai/laion-5b-a-new-era-of-open-large-scale-multi-modal-datasets/, 2022. 1

[4] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*, 2022. 1, 2, 3

[5] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 1, 2