NamedMask: Distilling Segmenters from Complementary Foundation Models

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

Abstract

The goal of this work is to segment and name regions of images without access to pixel-level labels during training. To tackle this task, we construct segmenters by distilling the complementary strengths of two foundation models. The first, CLIP [26], exhibits the ability to assign names to image content but lacks an accessible representation of object structure. The second, DINO [5], captures the spatial extent of objects but has no knowledge of object names. Our method, termed NamedMask, begins by using CLIP to construct category-specific archives of images. These images are pseudo-labelled with a category-agnostic salient object detector bootstrapped from DINO, then refined by category-specific segmenters using the CLIP archive labels. Thanks to the high quality of the refined masks, we show that a standard segmentation architecture trained on these archives with appropriate data augmentation achieves impressive semantic segmentation abilities for both single-object and multi-object images. As a result, our proposed NamedMask performs favourably against a range of prior work on five benchmarks including the VOC2012, COCO and large-scale ImageNet-S datasets.

1. Introduction

Semantic segmentation is a task that entails grouping and naming coherent regions of images. It has a broad range of applications spanning domains such as autonomous driving, manufacturing and medicine. A key barrier to automating this task through supervised learning is the requirement for pixel-level segmentation annotations, which can be extremely costly to obtain (e.g. 1.5 hours per image when accounting for quality control [9]).

The emerging paradigm of *foundation models* (models that have been pre-trained on broad data and can be adapted to a wide range of downstream tasks) has yielded striking gains for many machine perception problem domains [3]. There is therefore considerable interest in determining whether such models can alleviate the prohibitive annotation costs associated with semantic segmentation. In this vein, MaskCLIP [41] demonstrated the potential benefits of leveraging the representation learned by CLIP [26] to perform "annotation-free"[1] segmentation with no prior knowledge of the target domain. However, unless segmentation masks are available from at least some of the categories to guide pseudo-labelling, segmentation quality remains far from supervised performance. ReCo [30] considered an alternative formulation in which only the *names* of the target categories (but no images from the target domain) are available during training. By coupling the retrieval capabilities of CLIP to curate archives of images belonging to specific categories with a co-segmentation algorithm, ReCo obtained improvements over MaskCLIP but still struggles

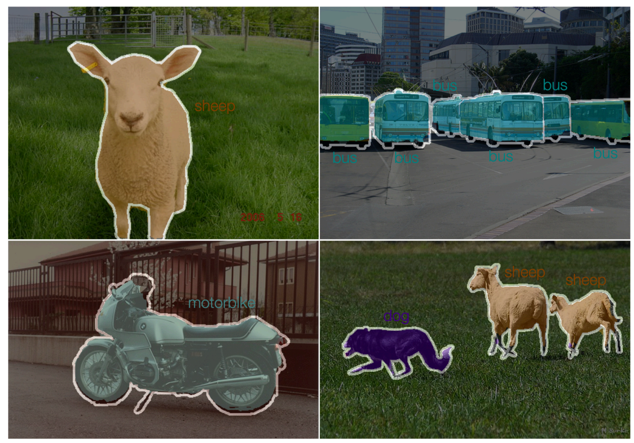

Figure 1. We propose NamedMask, a segmenter distilled from the complementary capabilities of CLIP and DINO. *With no access to pixel-level annotation*, NamedMask not only accurately segments single objects (left) but also multiple objects (right) within an image. White pixels denote ignored regions. The images are drawn from the VOC2012 benchmark.

[1]This setting can be equivalently referred to as *zero-shot transfer* in the terminology of [26].

to produce precise object segmentations.

In this work, we build upon the ReCo framework and revisit the mechanism through which it obtains pseudo-masks with semantic labels. Drawing inspiration from recent work showing that DINO features can be employed to perform unsupervised salient object detection [22, 29, 36], our first step is to replace the fragile co-segmentation of ReCo with a more robust category-agnostic object segmentation facilitated by DINO. We then exploit the naming capabilities of CLIP to assign the category label from each archive of images to these segmentations to enable the training of category-specific "expert" segmenters that refine the quality of the archive segmentations. Finally, we train a single semantic segmentation model on the resulting collection that is capable of segmenting objects from any category that is represented in the archives, using copy-paste augmentation [15] to improve generalisation to images of multiple objects. We show that our approach, which we term NamedMask, achieves substantial gains in performance for semantic segmentation of objects (see Fig. 1 for examples).

Our contributions are: (1) We propose NamedMask, a framework for segmenting and naming objects without pixel-level annotation by distilling the complementary strengths of CLIP and DINO; (2) We provide extensive experiments to demonstrate the improvements brought by NamedMask over prior semantic segmentation approaches that also make use of language-image pretraining.

2. Related work

Our approach relates to prior work on *unsupervised semantic segmentation*, *semantic segmentation with language-image pretraining* and *salient object detection*. We discuss connections to each of these next.

Unsupervised semantic segmentation. By coupling deep neural networks with creative learning objectives, substantial progress has been made towards unsupervised semantic segmentation. Examples of learning signals that have been constructed without labels include expectation-maximisation over segments [19], mutual information maximisation [20, 24], contrasting proposals [33], complementary signals from LiDAR and vision [34] and feature correspondence distillation [16]. In contrast to name-only segmentation, these methods do not make use of language-image pretraining or require access to the target category list during training. They do, however, require the use of a small number of images labelled with segments (typically the test set itself) together with the Hungarian algorithm to assign names to segment predictions, or otherwise employ nearest-neighbour lookup on a held-out set of images with segmentation masks.

Annotation-free semantic segmentation using language-image model. Several recent works have sought to leverage the zero-shot transfer capabilities of CLIP [26] to perform semantic segmentation with no access to paired data (images labelled with categories or segments) from the target domain. MaskCLIP [41] illustrated the potential of using CLIP for semantic segmentation in a zero-shot transfer setting (a setting that they term "annotation-free"). A recent example of such line of research is ReCo [30], which curates unlabelled images into examples of concepts with CLIP, then applies a co-segmentation algorithm to derive semantic segmentation training data. While ReCo achieves promising results, it fails to coherently pseudo-label objects and thus (as we show through experiments) does not lead to high-quality object segmentations. In this work, we compare directly with ReCo and demonstrate the substantial gains in performance that can be attained by bootstrapping the category-agnostic pseudo-labels enabled by DINO.

Unsupervised salient object detection. A range of work has sought to perform salient object detection (the task of segmenting foreground objects) without human annotation [2, 35, 39]. One notable trend amongst recent approaches has been the application of spectral clustering in combination with self-supervised features [22, 29, 36]. In this work, we build on the SelfMask approach of [29] to provide a robust category-agnostic segmenter for NamedMask.

3. Method

In this section, we formulate the semantic segmentation task considered in this work (Sec. 3.1) and the method, NamedMask, that we propose to tackle it (Sec. 3.2).

3.1. Task formulation and terminology

Our objective is to perform *semantic segmentation*: for a given image, $x \in \mathbb{R}^{3 \times H \times W}$, we aim to assign a label, c, from among a finite set of categories, $\mathcal{C}$, to each pixel location $\omega \in \{1, \dots, H\} \times \{1, \dots, W\}$ of x. To facilitate cost-effective scaling, we aim to do so without access to any form of pixel-level annotation. To this end, we propose to exploit the perceptual grouping of objects and their semantic categorisation offered by two foundation models. Specifically, we leverage the semantic categorisation capabilities of CLIP derived through large-scale language-image pretraining and the perceptual grouping capabilities of DINO derived from vision-only pretraining.

Terminology. To date, a wide array of methods have been proposed to tackle the problem of semantic segmentation with different levels of supervision (*fully unsupervised*, *unsupervised but with supervised pretraining*, *weakly-supervised* etc.). However, the terminologies used to describe these levels of supervision are not always clear or consistent. We therefore first aim to clarify the annotation regime in which we operate and how it is closely related to

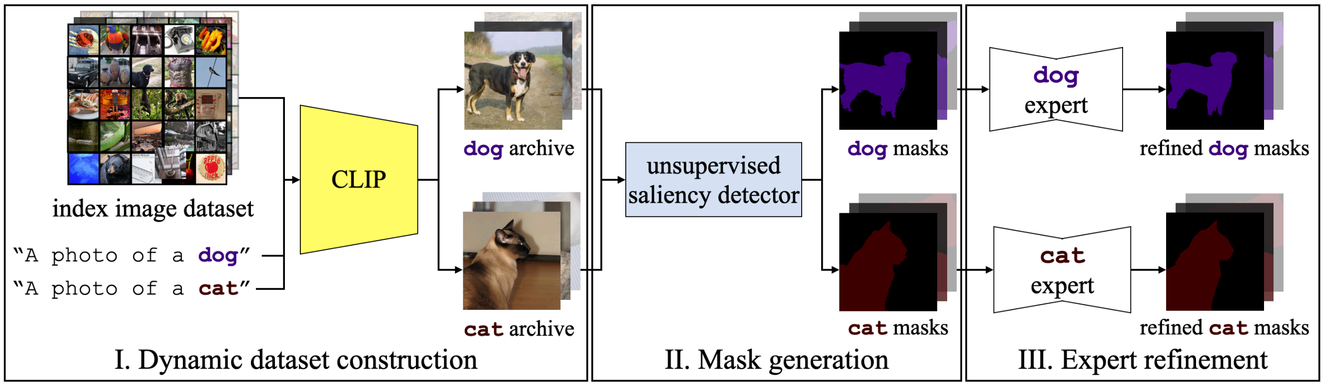

Figure 2. Overview of the proposed pipeline used to construct the NamedMask training dataset for semantic segmentation. Given an image archive for a concept retrieved by CLIP (left), we generate masks using an unsupervised saliency detector (middle). We refine the segmentations of each category by a class expert trained with the constructed image-mask pairs (right). Using the retrieved images and their refined segments, we train NamedMask to generate a segmenter capable of predicting a set of pre-defined categories (omitted in the figure for simplicity).

prior work.

In particular, we consider a setting that we term *Segmentation Leveraging Only Weak Pretraining* (SLOWP). SLOWP methods make no use of pixel-level annotation and are characterised by pretraining on data that is either: (1) *weakly-labelled* (e.g. with class labels or alt-text) and does not derive from the target distribution; or (2) *unlabelled* and may or may not derive from the target distribution. Within the space of SLOWP methods, we further distinguish three sub-categories that more precisely characterise the knowledge that the method possesses about the segmentation task used to evaluate the model: (i) *Zero-shot transfer* assumes no knowledge of the target distribution (images or category names) during training; (ii) *Name-only transfer* assumes access (during training) to the list of category names that are to be used for the target segmentation task, but does not assume access to any images from the target distribution; (iii) *Name-and-image transfer* assumes access (during training) to the list of category names in the target segmentation task *and* access to unlabelled images from the target distribution.

To relate these categories to prior work, note that MaskCLIP [41] represents a *SLOWP zero-shot transfer* method: it uses language-image pretraining (via CLIP) and does not make use of the target categories during training. ReCo [30] typically represents a *SLOWP name-only transfer* method: it uses language-image pretraining (via CLIP) and image classification pretraining (via DeiT-S/SIN [23]) and has access to target category names for constructing classifiers.

In this work, we propose a method, NamedMask, that operates effectively in both the *SLOWP name-only transfer* and *SLOWP name-and-image transfer* scenarios. We describe NamedMask next.

3.2. NamedMask

NamedMask is trained in a sequence of four stages: (1) For a given list of target categories, we perform dynamic dataset construction by curating archives of images for each category from an unlabelled image collection using CLIP; (2) For each image in each archive, we predict a category-agnostic object mask with an unsupervised saliency detector; (3) We refine the predicted masks with a category-specific "expert" segmenter, which is self-trained with the generated image-mask pairs within each archive; (4) We distill a segmenter using the image archives and their refined masks as pseudo-labels. An overview of the first three stages is provided in Fig. 2, and each stage is detailed in the following.

Dynamic archive construction. To create a data set containing images of categories of interest, we follow the approach proposed by ReCo [30] and curate an archive of images for each concept using an image-language model (*i.e.* CLIP). Formally, given an image encoder $\phi_{\mathcal{I}}$ and a text encoder $\phi_{\mathcal{T}}$ of CLIP, we curate one archive from an unlabelled image collection $\mathcal{U}$ for each category c of interest. We do so by selecting the n images among the collection whose visual embeddings $\phi_{\mathcal{I}}(x_i) \in \mathbb{R}^e$ lie closest to the text embedding[2] $\phi_{\mathcal{T}}(c) \in \mathbb{R}^e$ of c. That is,

$$\mathcal{A}_c = \{x_i \mid i \in \text{arg top}k\,[\phi_{\mathcal{I}}(\mathcal{U}) \cdot \phi_{\mathcal{T}}(c)]\}, \quad (1)$$

where $\mathcal{A}_c$ denotes the image archive for category c and arg top$k$ returns the indices of its arguments with the k largest values. In this way, we dynamically construct a data set comprising a collection of $|\mathcal{C}|$ archives (one for each category).

Mask generation. To produce category-agnostic object

[2]Details of the prompt used to construct the text embedding can be found in the supplementary material.

segmentations for the images within each archive, we adopt the SelfMask [29] unsupervised salient object detection method. SelfMask learns to perform salient object detection by first performing spectral clustering on DINO features across unlabelled images, then using these clusters as pseudo-labels to train a variant of MaskFormer segmenter [7]. Given the SelfMask saliency detector ψ_s, we first predict a category-agnostic saliency map $y_i = \psi_s(x_i) \in \{0,1\}^{H\times W}$ for each image $x_i \in \mathbb{R}^{3\times H\times W}$ in each archive. We then simply assign to each category-agnostic saliency map the category label c of the archive that contains the image. This produces a collection of images annotated with saliency masks and corresponding category labels.

Mask refinement through category experts. The category-agnostic saliency detector employed in the previous stage is unaware of the category of objects that it is being used to segment. We hypothesise that a segmenter will produce superior segmentations when it is given knowledge of the specific category of objects that it is required to segment, and thus will produce improved pseudo-masks for training NamedMask. To instantiate this idea, we refine the category-agnostic predictions made by the saliency detector with a segmenter ψ_c, which specialises in segmenting regions corresponding to category c. To this end, we train a segmenter ψ_c to assign regions to either the category c or the background class for each image in $\mathcal{A}_c$, as a pixel-level one-vs-all binary classification task. We then use the predictions obtained by ψ_c as pseudo-masks for category c. We show through experiments in Sec. 4.3 that this simple approach yields superior segmentation training data relative to using SelfMask predictions directly.

Note that for cases when there are a large number of target categories (*e.g.* 919 categories in ImageNet-S [13]), training one expert per class can be computationally expensive. For such cases, we group the relevant categories by applying k-means clustering to the text embeddings of the categories extracted from a CLIP text encoder and train an expert for each category group.

Training NamedMask. Given the resulting collection of image archives annotated by category-specific segmenters, NamedMask is produced by simply training a standard semantic segmentation architecture using a cross-entropy loss. Thus, self-training produces a segmenter that exploits the complementary information encoded by two different foundation models, where the visual-only model (*i.e.* DINO) implicitly captures the perceptual grouping of objects, and the ability to name categories derives from the language-image model (*i.e.* CLIP).

4. Experiments

In this section, we begin by describing the datasets considered for our experiments, implementation details, and our ablation study. We then compare our model to state-of-the-art unsupervised semantic segmentation (USS) methods and approaches that leverage only weak pretraining (SLOWP).

4.1. Datasets

Evaluation benchmarks. We consider five segmentation benchmarks including COCO [21], CoCA [40], Cityscapes [9], PASCAL VOC2012 [12], and ImageNet-S [13]. COCO consists of 118,287 and 5,000 images for train and validation splits with 80 object categories and a background class and CoCA comprises 1,295 images of 80 object categories with a background. Cityscapes contains 2,975 and 500 urban scene images for training and validation splits with 30 categories among which we pick 14 object categories to evaluate based on the original paper [9]. VOC2012 is composed of 1,464 training and 1,449 validation images with 21 categories including background, and the large-scale ImageNet-S [13] dataset comprises 9,190 train, 12,419 validation, and 27,423 test images with precise pixel-level annotations. There are three variations of ImageNet-S: ImageNet-S_{50}, ImageNet-S_{300}, and ImageNet-S_{919}, consisting of 50, 300, and 919 semantic categories of ImageNet1K [10], respectively.

We use the VOC2012 train split and the ImageNet-S_{300} validation split for our ablation studies, and compare our models to previous USS and SLOWP methods on CoCA, the validation split of COCO, Cityscapes, and VOC2012, and the test split of ImageNet-S.

Image collections. To curate image archives for each category, we use two unlabelled image collections: (1) For experiments on PASCAL VOC2012, we use the ImageNet1K training set without labels, following [30]. (2) For experiments on ImageNet-S, we use unlabelled images from LAION-5B [28]. For the latter, we implement the archive curation process using the CLIP feature index provided with the LAION-5B release[3]. Since the LAION-5B dataset was collected with limited manual curation, we apply a face detector to all images and discard any image containing a visible human face [11]. We refer the reader to the supplementary material for further details about the usage of the LAION-5B dataset.

4.2. Implementation details

We conduct the experiments on a single A100 NVIDIA graphic card with PyTorch [25]. Code will be made publicly available.

Network architecture and optimisation. We use DeepLabv3+ [6] with a ResNet50 [18] backbone for both category experts and NamedMask. We initialise the backbone with DINO [5] that is pretrained on ImageNet [27]

[3] https://laion.ai

model	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining-table	dog	horse	motorbike	person	potted plant	sheep	sofa	train	tv/monitor	avg.
SelfMask	78.5	26.8	68.0	59.2	22.9	69.3	43.1	80.7	14.0	70.2	20.3	74.2	70.1	57.7	41.5	20.6	74.3	18.4	64.7	25.0	50.0
experts	80.2	29.0	72.9	65.0	30.4	74.9	48.9	82.4	15.8	77.8	27.2	75.2	74.5	62.1	43.1	21.3	74.9	26.7	69.8	36.4	54.4

Table 1. **Category experts produce better quality segmentation masks than the baseline unsupervised saliency detector.** We report segmentation performance for each method on Pascal VOC2012. The performance metric is (class-wise) IoU (%).

in a self-supervised manner. For expert training, we adopt a lightweight learning schedule of 5K gradient updates with a batch size of 8 for COCO, CoCA, Cityscapes, and VOC2012 and 10K updates with a batch size of 16 for ImageNet-S. For NamedMask, we train the model with 20K iterations for COCO, CoCA, Cityscapes, and VOC2012 and 80K iterations for ImageNet-S. We use standard data augmentations such as random scaling, random cropping and colour jittering. We use Adam optimiser with an initial learning rate of 0.0005 and a weight decay of 0.0002. We decay the learning rate with the Poly learning rate [6].

To curate category archives from ImageNet and LAION-5B, the ViT-L/14@336px and ViT-L/14 variants of CLIP are employed repectively. For our unsupervised saliency detection method, we adopt the model from SelfMask [29], and apply a bilateral solver [1] to predictions from SelfMask as a post-processing step.

Inference. When evaluating on ImageNet-S, images are resized such that their larger dimension is 1024 pixels while preserving their aspect ratio. For evaluation, the predictions of the model are then resized back to the original resolution to match the ground-truth mask by using a bilinear upsampler. For the ImageNet-S_{300} validation set (used in our ablation study), we resize the shorter side of images to 384 with a maximum length for the larger dimension of 512 pixels. For the other benchmarks, we use the original resolution of the images.

Metric. Following the common practice, we employ intersection-over-union (IoU) to measure a class-agnostic mask quality and mean IoU (mIoU) to evaluate the performance of semantic segmentation.

4.3. Ablation study

In this section, we present a thorough ablation study on each component of our proposed NamedMask, namely, the influence of archive size, the influence of adopting category experts and the effect of the number of category experts. We also investigate the influence of adopting copy-paste augmentation for segmenting images with multiple objects.

Effect of archive size. Unlike supervised approaches for which it is costly to acquire annotations, the dataset creation process for NamedMask can be easily scaled. To investigate the influence of the number of images used for training, we vary the size of archive curated by CLIP for each category, from 1 to 500 images, and train NamedMask on the resulting images with corresponding pseudo-labels obtained from SelfMask. As for quantitative evaluation, we adopt the VOC2012 training set and report numbers in Fig. 3. Note that, the training is done only on the constructed archive of ImageNet images, with pseudo labels acquired from SelfMask (*i.e.* no category experts have been introduced at this stage).

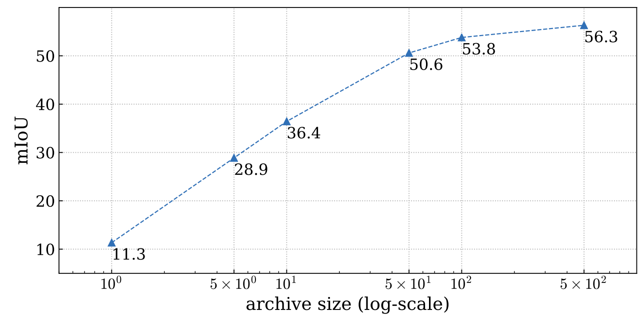

Figure 3. **Larger image archives produce better segmenters.** Here, "archive size" denotes the number of images retrieved by CLIP to curate an image archive for each category.

As shown in Fig. 3, that the archive size plays an important role in the performance of our model, monotonically increasing with the number of images for an archive. For the remaining experiments, we curate 500 images per archive.

Effect of category experts on mask quality. As described in Sec. 3.2, we propose to refine the pseudo-labels from SelfMask with category-specific experts, which are trained to distinguish foreground and background pixels.

To compare the quality of category-agnostic masks generated by SelfMask and class-specific masks by an expert, we evaluate their predictions on 20 object categories from the VOC2012 train split. Specifically, we train 20 category experts on image archives constructed by retrieving from ImageNet1K. In Tab. 1, we show the (class) IoU of each category. We observe that the experts consistently outperform SelfMask across all categories. For a qualitative comparison, we visualise the examples predictions from both SelfMask and category experts in Fig. 4. In the following experiments, we produce pseudo-masks from an expert for each category by default.

Training an expert for a category group. When there are numerous classes that are semantically close to one another, training individual category experts may become prohibitively expensive. We therefore group categories by ap-

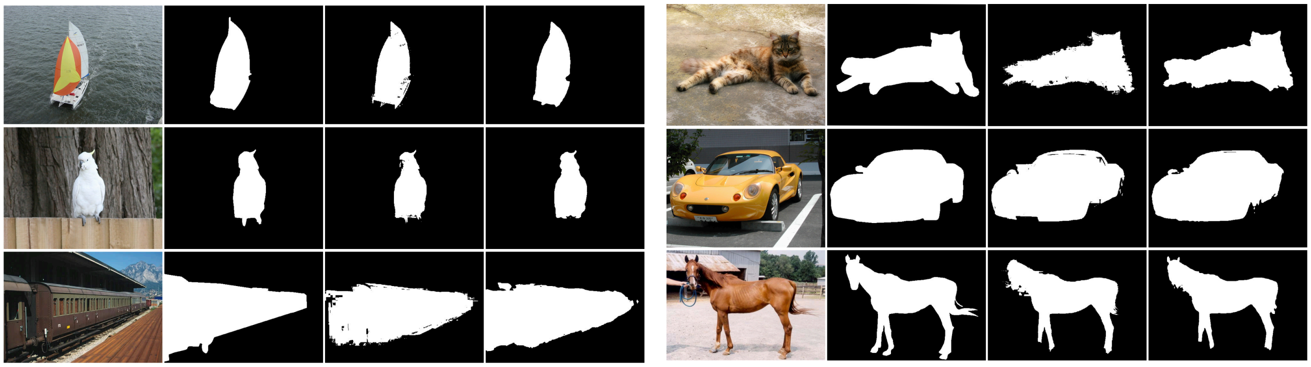

Figure 4. **Category experts refine the masks provided by an unsupervised saliency detector (*i.e.* SelfMask).** The images are selected from VOC2012. Zoom in for details.

model	# experts	avg. IoU
SelfMask [29]	-	62.7
category experts	1	63.3
	30	**64.1**
	60	64.0
	90	63.9

Table 2. Effect of grouping semantically relevant categories for category expert training on the ImageNet-S_{300} validation split.

model	copy-paste	single obj.	multi-obj.	all
SelfMask + CLIP	-	63.3	42.1	50.4
NamedMask (Ours)	✗	67.0	50.5	56.6
NamedMask (Ours)	✓	**68.0**	**53.6**	**58.7**

Table 3. **Copy-paste augmentation helps the model to segment multiple objects in an image.** The performance is measured in mIoU (%). A baseline model is marked in gray. Best score for each column is highlighted in **bold**.

plying k-means to text embeddings of the categories into k concept groups as described in Sec. 3.2.

To show the effect of k, we evaluate the category experts, by varying the number of category groups on the ImageNet-S_{300} validation set. In Tab. 2, we report average IoU over 300 categories for k={30, 60, 90} which corresponds to 10%, 20%, and 30% of the total number of classes, respectively. As baselines, we also show average IoU for SelfMask and a single expert for 300 categories (*i.e.* k=1). As can be seen, training an expert always shows higher score than the baseline, even when k is set to 1. However, when we set the number of experts higher than 30, the performance appears to saturate. We conjecture that this is because there is a trade-off between the number of different images which an expert can see during training and average semantic relevance within a category group. That is, when k is large, the average number of categories assigned to one group tends to be low, which decreases the total number of

model	dataset
MaskCLIP [41]	WebImageText
ReCo [30]	WebImageText, (Stylized-)ImageNet
NamedMask (Ours)	WebImageText, ImageNet

Table 4. Datasets employed in training each model. ReCo utilises both Stylized-ImageNet and ImageNet.

images used for training an expert for the group. In contrast, when k is small, the total number of images for an expert become large at the cost of reduced semantic relevance in a class group. For this reason, we group categories of ImageNet-S_{300} and -S_{919} into 30 and 90 which (approx.) accounts for 10% of total classes, respectively.

Effect of copy-paste on segmenting multiple objects. In contrast to the salient object detectors and category experts which segment an object of a single category or a group of similar categories within an image, our model can be readily trained to segment multiple objects of different categories by employing the copy-paste augmentation [15]. To demonstrate this, we evaluate two NamedMask models, trained with and without copy-paste augmentation on the images containing a single object or multi-objects from possibly different categories. As a baseline, we also evaluate segmentation of SelfMask whose semantic label is decided by applying CLIP to a given image. As shown in Tab. 3, when validated on the VOC2012 training split, copy-paste brings a notable gain in performance by 4.1 mIoU compared to the model trained without copy-paste. We therefore adopt the copy-paste augmentation as a default setting in the remaining experiments.

4.4. Comparison to state-of-the-art methods

To describe the effectiveness of our approach, we compare NamedMask against existing approaches that fall into the proposed segmentation leveraging only weak pretraining (SLOWP) setting. Specifically, we consider and re-implement MaskCLIP [41] with the *zero-shot transfer* set-

model	transfer type	COCO	CoCA	Cityscapes$_{obj}$
MaskCLIP [41]	zero-shot	5.3	3.1	6.1
ReCo† [30]	name-only	17.1	16.9	14.1
NamedMask	name-only	**28.4**	**27.3**	**18.2**

Table 5. Comparison to previous segmentation leveraging only weak pre-training (SLOWP) methods on the COCO, CoCA, and Cityscapes$_{obj}$ benchmarks in terms of mIoU. Highest scores on each benchmark are in **bold**. †initialises the backbone with Stylized-ImageNet pre-training.

model	transfer type	backbone	mIoU
USS			
Inst. Disc. [37]	-	ResNet50	4.3
MoCo [17]	-	ResNet50	3.7
InfoMin [31]	-	ResNet50	4.4
SwAV [4]	-	ResNet50	4.4
MaskCon. [32]	-	ResNet50†	35.0
MaskDist. [33]	-	ResNet50†	**45.8**
SLOWP			
MaskCLIP* [41]	zero-shot	ResNet50	29.1
ReCo*‡ [30]	name-only	DeiT-S/16	34.2
NamedMask	name-only	ResNet50	**59.2**

Table 6. Comparison to existing unsupervised semantic segmentation (USS) and segmentation leveraging only weak pre-training (SLOWP) methods on the PASCAL VOC2012 validation set. Numbers for *USS* methods are from MaskDistill. *Re-implemented and adapted by us to predict a background class. †uses dilated ResNet [38]. ‡initialises the backbone with Stylized-ImageNet [14] pre-training. Highest scores of each kind of methods are in **bold**.

ting (*i.e.* the annotation-free setting in their paper) and ReCo [30] with the *name-only transfer* or and *name-and-image transfer* setting. As described in Sec. 3.1, the transfer type of each SLOWP method is determined by whether it has access to either category names or unlabelled images from the evaluation benchmark during training. As such, the transfer type of a method varies with its evaluation benchmark (see Tab. 4 for datasets of which categories and images each SLOWP approach has access to during training).

As MaskCLIP and ReCo do not explicitly define background categories, we classify the pixels as background if their highest class probability is lower than a certain threshold t. We set t as 0.25 and 0.9 for MaskCLIP and ReCo (see the supp. mat. for more details on how t is selected).

We evaluate on popular segmentation benchmarks including COCO, CoCA, VOC2012, and large-scale ImageNet-S datasets. Additionally, we also evaluate on the object categories (*e.g.* car, person) in Cityscapes (denoted Cityscapes$_{obj}$). In addition to SLOWP methods, we also compare with state-of-the-art unsupervised semantic segmentation (USS) including MaskContrast [32] and MaskDistill [33] on the VOC2012 and ImageNet-S benchmarks, as they share with SLOWP the similar goal of training without manual annotations.

In Tab. 5, we compare NamedMask to previous SLOWP methods on COCO, CoCA, and Cityscapes$_{obj}$ benchmarks. We make two observations: (i) ReCo and NamedMask, which have access to the category names, outperform MaskCLIP, which is unaware of the concepts of the target benchmarks during training; (ii) when comparing the two name-only transfer methods, NamedMask performs better than ReCo by a large margin on each dataset.

In Tab. 6, we report the results of NamedMask on the VOC2012 validation split. Our approach shows favourable performance over the existing models for both SLOWP and USS. In detail, while the previous SLOWP methods fall behind the state-of-the-art USS models, NamedMask outperforms them by some (≈13.4 mIoU). We also observe that the proposed method is competitive on ImageNet-S, which consists of significantly more number of categories than VOC2012. Here, NamedMask corresponds to the *name-and-image transfer* setting since it has implicit access to unlabelled images from the target distribution through its use of SelfMask (which is bootstrapped from DINO). Similarly, ReCo is categorised as *name-and-image transfer*, as it uses ImageNet1K training images for constructing classifiers. With the caveat that each method has access to different information, NamedMask outperforms the state-of-the-art methods by 15.5, 14.7, and 11.9 mIoU on ImageNet-S$_{50}$ (in Tab. 7), ImageNet-S$_{300}$ (in Tab. 8), and ImageNet-S$_{919}$ (in Tab. 9), respectively.

For qualitative results, we show sample visualisations of our method in Fig. 1. More visualisation examples including failure cares are shown in the supplementary material.

5. Limitations

We note several limitations of our approach: (1) We need to train a new segmenter each time we wish to include another category which is not considered in the previous training of NamedMask. Future work could potentially address this by developing a segmenter that directly predicts embeddings in the shared textual embedding space of CLIP. These could subsequently be used for naming predictions beyond the categories seen during training (i.e. generalisation to unseen categories without retraining). (2) While we primarily focus on object semantic segmentation by leveraging an unsupervised saliency detector, it would strengthen our approach to incorporate cues to segment "stuff" categories such as water, sky, etc. This could potentially be done by building prior work such as ReCo or MaskCLIP, which are capable of predicting stuff categories, into our pseudo-label generation step. (3) We note that NamedMask struggles to

model	transfer type	mIoU	S	MS	ML	L
USS						
MDC [8]	-	3.6	0.4	2.6	3.8	4.9
MDC† [8]	-	14.3	2.6	10.9	14.6	19.1
PiCIE [8]	-	4.5	0.2	3.1	5.0	5.3
PiCIE† [8]	-	17.6	4.4	13.1	20.1	23.1
MaskCon. [32]	-	24.2	**12.2**	25.6	24.7	20.4
LUSS$_s$ [13]	-	29.3	6.6	25.0	33.2	32.6
LUSS$_p$ [13]	-	**32.0**	9.7	**26.2**	**36.5**	**40.5**
SLOWP						
MaskCLIP* [41]	zero-shot	17.9	3.6	13.1	18.6	20.6
ReCo*† [30]	name & image	22.6	10.0	24.6	22.1	18.8
NamedMask	name & image	**47.5**	**23.5**	**48.7**	**49.3**	**38.0**

Table 7. Comparison to existing *USS* and *SLOWP* methods on the ImageNet-S_{50} benchmark. Scores for *USS* are drawn from LUSS [13]. We also report mIoU under different object sizes from small (S), medium-small (MS), medium-large (ML), and large (L). *Re-implemented and adapted by us to predict a background class. †initialises the encoder with supervised ImageNet (for MDC and PiCIE) or Stylized-ImageNet pre-training (for ReCo). Best score for each column within a same method type is in **bold**.

model	transfer type	mIoU	S	MS	ML	L
USS						
LUSS$_s$ [13]	-	16.0	2.8	12.0	16.4	21.7
LUSS$_p$ [13]	-	**18.1**	**4.2**	**13.6**	**19.5**	**23.5**
SLOWP						
MaskCLIP* [41]	zero-shot	1.6	0.4	0.6	1.2	2.5
ReCo*† [30]	name & image	8.5	5.4	9.7	8.4	5.6
NamedMask	name & image	**32.8**	**9.9**	**29.1**	**34.9**	**26.0**

Table 8. Evaluation on the ImageNet-S_{300} benchmark. *Re-implemented and adapted by us to predict a background class. †initialises the encoder with Stylized-ImageNet pre-training.

differentiate a category which often appears with another concept. For example, a rider of a motorbike sometimes is classified as part of the motorbike. We conjecture that this is due to the use of a (unsupervised) saliency detector for pseudo-mask generation which highlights dominant regions in an image without account for a semantic of interest. We believe this can be alleviated by considering a semantic prediction for an expected category in an image (similary to the language-guided attention mechanism used in ReCo [30]) in addition to the prediction made by the saliency detector.

6. Broader impact

NamedMask distills segmenters from foundation models. While powerful, these models have been shown to exhibit biases across different racial and religious groups [3]. It is therefore likely that NamedMask inherits these biases to some degree. As such NamedMask represents a research prototype that is not appropriate for real-world usage without additional consideration of the deployment setting and the design of appropriate mitigation mechanisms.

model	transfer type	mIoU	S	MS	ML	L
USS						
LUSS$_s$ [13]	-	6.6	1.3	4.6	7.1	8.4
LUSS$_p$ [13]	-	**11.0**	**2.4**	**8.3**	**11.9**	**13.4**
SLOWP						
MaskCLIP* [41]	zero-shot	0.5	0.1	0.2	0.3	0.8
ReCo*† [30]	name & image	3.8	2.6	4.6	3.6	2.5
NamedMask	name & image	**22.9**	**5.1**	**19.4**	**24.4**	**19.8**

Table 9. Evaluation on the ImageNet-S_{919} benchmark. *Re-implemented and adapted by us to predict a background class. †initialises the encoder with Stylized-ImageNet pre-training. NamedMask is able to segment reasonably well even when numerous categories (*i.e.* 919 classes) are present in the target dataset.

NamedMask aims to achieve semantic segmentation with a methodology that can be scaled up without the prohibitive cost of manually-collected segment annotation. In doing so, we hope that it will help enable the deployment of semantic segmentation for applications that yield positive societal impact. As with many powerful computer vision technologies, however, NamedMask is a tool that is subject to *dual use* and is therefore vulnerable to abuse. We are likely unable to anticipate all such possible abuses, but examples could include applications that entail unlawful surveillance.

7. Conclusion

In this work, we introduced NamedMask, a method for semantic segmentation that is trained by distilling the complementary capabilities of two foundation models, CLIP and DINO, into a single segmenter. By doing so, NamedMask achieves impressive segmentation quality across both single-object and multi-object images without pixel-level annotation. We demonstrate the effectiveness of NamedMask by comparing to prior methods on several standard semantic segmentation benchmarks including the large-scale ImageNet-S_{919} dataset, where we observe that NamedMask achieves a significant boost in segmentation performance.

Acknowledgements. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). GS is supported by AI Factory, Inc. in Korea. GS would like to thank Guanqi Zhan for proof-reading. SA would like to acknowledge the support of Z. Novak and N. Novak in enabling his contribution.

References

[1] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *ECCV*, 2016. 5

[2] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *NeurIPS*, 2019. 2

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021. 1, 8

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 7

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 4

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 4, 5

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 4

[8] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, 2021. 8

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 4

[12] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015. 4

[13] 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. 4, 8

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 7

[15] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 2, 6

[16] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[19] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, 2019. 2

[20] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4

[22] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. *arXiv:2205.07839*, 2022. 2

[23] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021. 3

[24] Yassine Ouali, Céline Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *ECCV*, 2020. 2

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 4

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 4

[28] 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. 4

[29] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *CVPRW*, 2022. 2, 4, 5, 6

[30] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*, 2022. 1, 2, 3, 4, 6, 7, 8

[31] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020. 7

[32] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, 2021. 7, 8

[33] Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. *arXiv:2206.06363*, 2022. 2, 7

[34] Antonin Vobecky, David Hurych, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. *arXiv:2203.11160*, 2022. 2

[35] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *ICML*, 2020. 2

[36] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *CVPR*, 2022. 2

[37] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 7

[38] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 7

[39] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *ICCV*, 2017. 2

[40] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *ECCV*, 2020. 4

[41] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 1, 2, 3, 6, 7, 8