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 do-

---

3This 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 mechanisim 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, \( C \), to each pixel location \( \omega \in \{1, \ldots, H\} \times \{1, \ldots, W\} \) 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
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 $\phi_I$ and a text encoder $\phi_T$ of CLIP, we curate one archive from an unlabelled image collection $U$ for each category $c$. We do so by selecting the $n$ images among the collection whose visual embeddings $\phi_I(x_i) \in \mathbb{R}^e$ lie closest to the text embedding $\phi_T(c) \in \mathbb{R}^e$ of $c$. That is,

$$A_c = \{x_i \mid i \in \arg \max_k [\phi_T(U) \cdot \phi_T(c)]\},$$

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

Mask generation. To produce category-agnostic object

---

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).
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 $\psi_s$, 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 $\psi_c$, which specialises in segmenting regions corresponding to category $c$. To this end, we train a segmenter $\psi_c$ to assign regions to either the category $c$ or the background class for each image in $A_c$, as a pixel-level one-vs-all binary classification task. We then use the predictions obtained by $\psi_c$, 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$_{300}$, ImageNet-S$_{3000}$, 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
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-S300 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).

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

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

<table>
<thead>
<tr>
<th>model</th>
<th># experts</th>
<th>avg. IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SelfMask [29]</td>
<td>-</td>
<td>62.7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>63.3</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td><strong>64.1</strong></td>
</tr>
<tr>
<td>category experts</td>
<td>60</td>
<td>64.0</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>63.9</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>model</th>
<th>copy-paste</th>
<th>single obj.</th>
<th>multi-obj.</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>SelfMask + CLIP</td>
<td>-</td>
<td>63.3</td>
<td>42.1</td>
<td>50.4</td>
</tr>
<tr>
<td>NamedMask (Ours)</td>
<td>✓</td>
<td><strong>67.0</strong></td>
<td><strong>50.5</strong></td>
<td>56.6</td>
</tr>
<tr>
<td>NamedMask (Ours)</td>
<td>✓</td>
<td>68.0</td>
<td><strong>53.6</strong></td>
<td><strong>58.7</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>model</th>
<th>dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskCLIP [41]</td>
<td>WebImageText</td>
</tr>
<tr>
<td>ReCo [30]</td>
<td>WebImageText, (Stylized-)ImageNet</td>
</tr>
<tr>
<td>NamedMask (Ours)</td>
<td>WebImageText, ImageNet</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>model</th>
<th>transfer type</th>
<th>COCO</th>
<th>CoCA</th>
<th>Cityscapes\textsubscript{obj}</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskCLIP [41]</td>
<td>zero-shot</td>
<td>5.3</td>
<td>3.1</td>
<td>6.1</td>
</tr>
<tr>
<td>ReCo [30]</td>
<td>name-only</td>
<td>17.1</td>
<td>16.9</td>
<td>14.1</td>
</tr>
<tr>
<td>NamedMask</td>
<td>name-only</td>
<td><strong>28.4</strong></td>
<td><strong>27.3</strong></td>
<td><strong>18.2</strong></td>
</tr>
</tbody>
</table>

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. †initialises the backbone with Stylized-ImageNet \[14\] pre-training. Highest scores of each kind of methods are in **bold**.

<table>
<thead>
<tr>
<th>model</th>
<th>transfer type</th>
<th>backbone</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>USS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inst. Disc. [37]</td>
<td>-</td>
<td>ResNet50</td>
<td>4.3</td>
</tr>
<tr>
<td>MoCo [17]</td>
<td>-</td>
<td>ResNet50</td>
<td>3.7</td>
</tr>
<tr>
<td>InfoMin [31]</td>
<td>-</td>
<td>ResNet50</td>
<td>4.4</td>
</tr>
<tr>
<td>SwAV [4]</td>
<td>-</td>
<td>ResNet50</td>
<td>4.4</td>
</tr>
<tr>
<td>MaskCon. [32]</td>
<td>-</td>
<td>ResNet50[1]</td>
<td>35.0</td>
</tr>
<tr>
<td>MaskDist. [33]</td>
<td>-</td>
<td>ResNet50[1]</td>
<td><strong>45.8</strong></td>
</tr>
</tbody>
</table>

**SLOWP**

<table>
<thead>
<tr>
<th>model</th>
<th>transfer type</th>
<th>backbone</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MaskCLIP* [41]</td>
<td>zero-shot</td>
<td>ResNet50</td>
<td>29.1</td>
</tr>
<tr>
<td>ReCo*† [30]</td>
<td>name-only</td>
<td>DeiT-S/16</td>
<td>34.2</td>
</tr>
<tr>
<td>NamedMask</td>
<td>name-only</td>
<td>ResNet50</td>
<td><strong>59.2</strong></td>
</tr>
</tbody>
</table>

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\[919\] (in Tab. 8), and ImageNet-S\[300\] (in Tab. 9), respectively.

For qualitative results, we show sample visualisations of our method in Fig. 1. More visualisation examples including failure cases 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
to some degree. As such, NamedMask represents a research endeavor that might 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 with additional consideration of the deployment setting and the design of appropriate mitigation mechanisms.

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-S319 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


[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