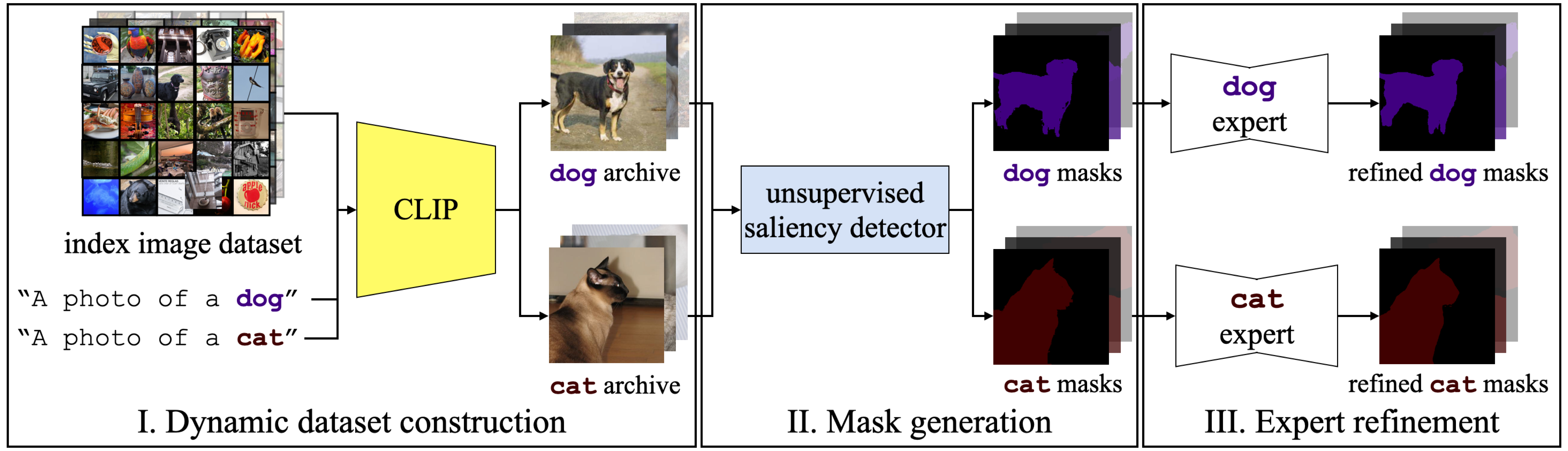
# NamedMask:
## Distilling Segmenters from Complementary Foundation Models

**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

## Overview



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).

## Contributions

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

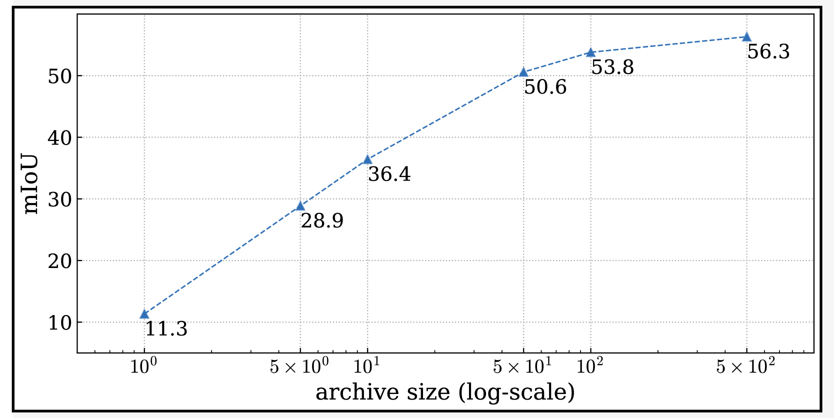
## Task formulation and terminology

We consider a setting for segmentation 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:

(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.

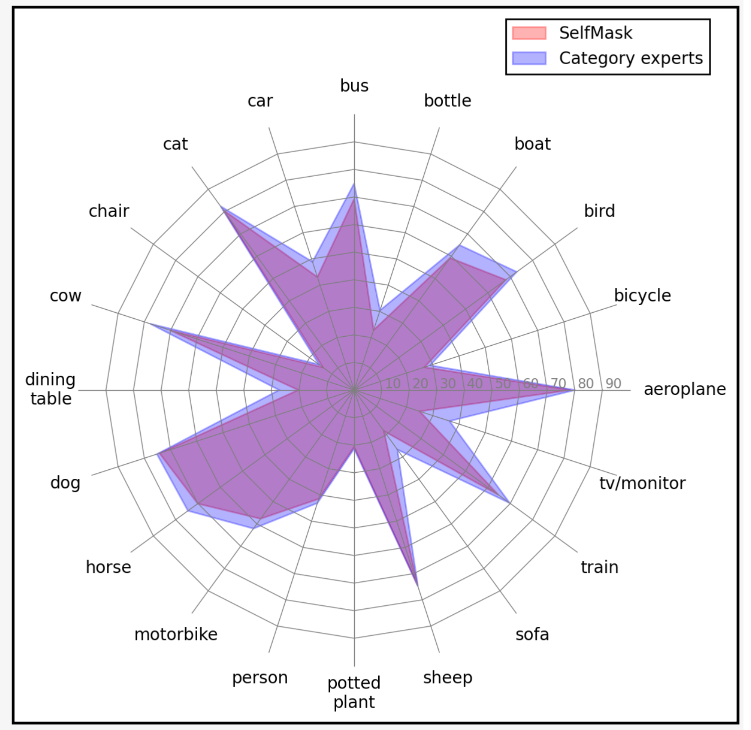
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

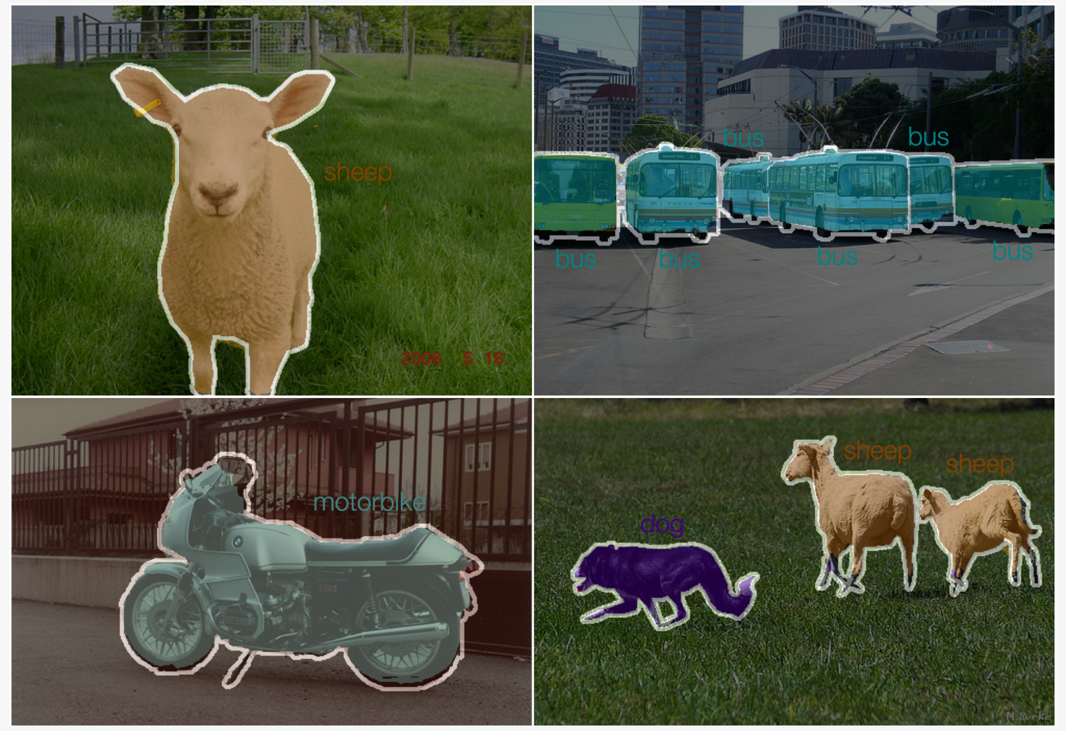
## Ablation studies



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

**Effect of category experts on mask quality.** 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 (%).





**Sample visualisations of NamedMask, a segmenter distilled from the complementary capabilities of CLIP and DINO.**

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

**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** |

**Copy-paste augmentation helps the model to segment multiple objects in an image.** The performance is measured in mIoU (%).

## Main results

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

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

| model | transfer type | backbone | mIoU |
|---|---|---|---|
| *unsupervised semantic segmentation methods* | | | |
| Inst. Disc. | - | ResNet50 | 4.3 |
| MoCo | - | ResNet50 | 3.7 |
| InfoMin | - | ResNet50 | 4.4 |
| SwAV | - | ResNet50 | 4.4 |
| MaskCon. | - | ResNet50† | 35.0 |
| MaskDist. | - | ResNet50† | **45.8** |
| *segmentation leveraging only weak pretraining methods* | | | |
| MaskCLIP* | zero-shot | ResNet50 | 29.1 |
| ReCo*‡ | name-only | DeiT-S/16 | 34.2 |
| NamedMask | name-only | ResNet50 | **59.2** |

**Comparison to existing unsupervised semantic segmentation and segmentation leveraging only weak pretraining methods on the PASCAL VOC2012 validation set.** *Re-implemented and adapted by us to predict a background class. †uses dilated ResNet. ‡initialises the backbone with Stylized-ImageNet pretraining.

## 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 benchmarks, where we observe that NamedMask achieves a significant boost in segmentation performance.