# Zero-shot Unsupervised Transfer Instance Segmentation

**Gyungin Shin**[1, 2] **Samuel Albanie**[2] **Weidi Xie**[1, 3]

[1]Visual Geometry Group, University of Oxford, UK

[2]Cambridge Applied Machine Learning Lab, University of Cambridge, UK

[3]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, China

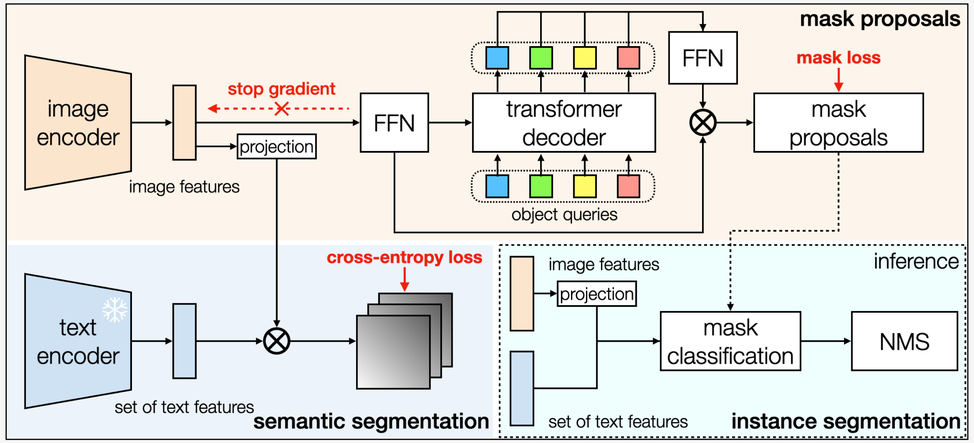
## Contributions

- We introduce a challenging task, namely, zero-shot unsupervised transfer instance segmentation, which aims to segment object instances without human supervision or access to a target data distribution.
- We propose a simple yet effective framework, termed ZUTIS, that goes beyond prior USSLIP approaches, and enables to concurrently perform instance segmentation in addition to semantic segmentation.
- We show that ZUTIS performs favourably against state-of-the-art methods on standard unsupervised segmentation benchmarks (e.g., COCO, ImageNet-S) by a large margin in both zero-shot transfer and unsupervised domain adaptation settings.

## Qualitative examples

**Sample visualisations of ZUTIS on COCO-20K and VOC2012.**

## Overview




**Step 0. (Before training)** We generate pseudo-masks by applying an unsupervised saliency detector (i.e., SelfMask) to images curated with CLIP for a set of categories of interest. For simplicity, this step is omitted in the figure.

**Step 1. (Training)** We feed an image to a CLIP image encoder whose resulting image features are given to a feed-forward network (FFN) followed by a transformer decoder to produce mask proposals which are updated through a mask loss **(top)**. At the same time, the CLIP image features are projected into a text embedding space in which semantic predictions are made via a dot-product between the projected image features and frozen text features for a set of categories **(bottom left)**. The semantic predictions are guided by the standard cross-entropy loss.

**Step 2. (Inference)** We predict instance segmentation masks using both the objectness score and the classification score of a mask proposal, after which we apply non-maximum suppression **(bottom right)**. For semantic segmentation, we follow the same process as during the training step, computing a dot-product between the projected image features and the frozen text features corresponding to a set of categories.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

## Main results

| model | backbone | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|---|
| *unsupervised methods w/o language-image pretraining* | | | | |
| DINO | ViT-S/16 | 0.7 | 2.0 | 0.4 |
| LOST | ViT-S/16 | 1.2 | 3.3 | 0.6 |
| MaskDistill | ViT-S/16 | 1.7 | 4.1 | 1.4 |
| MaskDistill† | RN50-C4 | 3.5 | 7.7 | 2.9 |
| *unsupervised methods w/ language-image pretraining* | | | | |
| MaskCLIP | ViT-B/32 | 0.3 | 0.8 | 0.2 |
| ZUTIS (Ours) | ViT-B/32 | 3.4 | 8.0 | 2.6 |
| MaskCLIP | ViT-B/16 | 1.3 | 3.4 | 0.8 |
| ZUTIS (Ours) | ViT-B/16 | **5.7** | **11.0** | **5.4** |

**Comparison to previous unsupervised instance segmentation methods on COCO-20K.** †Mask R-CNN trained with pseudo-masks from MaskDistill.

| model | arch. | COCO | CoCA |
|---|---|---|---|
| *initialised with different encoder features* | | | |
| ReCo† | DeiT-S/16 & RN50x16 | 23.8 | 28.8 |
| NamedMask‡ | RN50 & DLv3+ | 28.4 | 27.3 |
| *initialised with CLIP encoder features* | | | |
| MaskCLIP | ViT-B/16 | 20.6 | 20.2 |
| ZUTIS (Ours) | ViT-B/16 | **32.8** | **32.7** |

**Comparison to previous unsupervised semantic segmentation methods leveraging image-language pretraining on COCO and CoCA in terms of mIoU (%).** †Initialised with supervised Stylised-ImageNet pretraining. ‡Initialised with DINO.

| model | arch. | # params | mIoU |
|---|---|---|---|
| *unsupervised methods w/o language-image pretraining* | | | |
| $PASS_p$ | RN50 | 25.6 | 6.6 |
| $PASS_s$ | RN50 | 25.6 | 11.0 |
| *unsupervised methods w/ language-image pretraining* | | | |
| ReCo† | DeiT-S/16 & RN50x16 | 170.4 | 10.3 |
| NamedMask‡ | RN50 & DLv3+ | 26.6 | 22.9 |
| ZUTIS (Ours) | ViT-B/32 | 87.8 | 27.5 |
| ZUTIS (Ours) | ViT-B/16 | 86.2 | **37.4** |

**Comparison to existing unsupervised methods on the ImageNet-S benchmark with 919 object categories in the unsupervised domain adaptation setting.**

| category-specific label | CUB-200-2011 |
|---|---|
| ✗ | 72.5 |
| ✓ | 72.6 |

**High-level to low-level zero-shot transfer on the CUB-200-2011 benchmark.** When given a fine-grained bird breed, ZUTIS can segment the corresponding bird regions as good as when it is given a high-level category "bird."

| model | AP | $AP_{50}$ | $AP_{75}$ |
|---|---|---|---|
| MaskCLIP | 0.7 | 2.0 | 0.4 |
| ZUTIS (Ours) | **3.3** | **7.2** | **2.8** |

**Zero-shot unsupervised instance segmentation for 15 unseen categories on COCO-20K**

## Conclusion

In this work, we introduced ZUTIS, the first framework for joint instance segmentation and semantic segmentation in a zero-shot transfer setting that requires no pixel-level or instance-level annotation. We employ a query-based transformer architecture for instance segmentation and train it on pseudo-labels generated from applying an unsupervised saliency detector to images retrieved by CLIP. Through careful experiments, we demonstrated the effectiveness of ZUTIS across both instance segmentation and semantic segmentation tasks.