Zero-shot Unsupervised Transfer Instance Segmentation Supplementary Material

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
https://www.robots.ox.ac.uk/~vgg/research/zutis

In this supplementary material, we provide a pseudo-code for ZUTIS (in Sec. A) and further details about our experiments (in Sec. B). Then, we describe additional ablation studies with regards to copy-paste augmentation [5] and a hyperparamter choice for temperature used to compute a mask confindence score (in Sec. C). Lastly, we show additional visualisations including common failure cases (in Sec. D).

A. Pseudo-code

In Algorithm 1, we describe a pseudo-code for a forward pass of ZUTIS. For readability, we omit operations for bilinear upsampling, which is applied to image features from an image encoder, and non-maximum suppression, which is applied to mask proposals.

B. Experiment details

Here, we detail the network architectures used for ReCo [10] and NamedMask [11] and their differences from ZUTIS. Next, we describe the details of data augmentations used to train our model.

B.1. Architecture details for ReCo and NamedMask

In Sec. 4.3, we compare ZUTIS to previous methods for unsupervised semantic segmentation with language-image pre-training. Here, we describe in more detail the differences between ZUTIS and ReCo [10], as well as the concurrent work NamedMask [11].

ReCo. ReCo is composed of two different image encoders and a text encoder. For the former, it adopts DeiT-S/16 [7] pretrained on Stylised-ImageNet [4] in a supervised manner and ResNet50x16 from CLIP [8], which is used for language-guided co-segmentation [10] together with the text encoder. Unlike ReCo, ZUTIS involves a single image encoder from CLIP (ViT-B/32 or ViT-B/16) and a corresponding text encoder streamlining the inference process.

NamedMask. NamedMask consists of an image encoder, which is ResNet50 [6] pretrained on the ImageNet1K dataset in an unsupervised manner [1], and an image decoder, which is DeepLabv3+ [2]. It is worth noting that NamedMask follows a conventional semantic segmentation architecture which can only predict a pre-fixed set of classes that the model has seen during training. In contrast, due to the use of a text encoder as a classifier, ZUTIS can readily predicts categories beyond seen ones after training for a set of concepts.

B.2. Data augmentation

For geometric transformations, we use random horizontal flipping with a probability of 0.5, random rescaling with a range of [0.1, 1.0], and random cropping with a size of 384×384. For photometric transformations, we use random colour jittering and gray scaling with a probability of 0.8 and 0.2, respectively. We also use a random gaussian blurring with a kernel size of 10% of a shorter side of an image. Lastly, we apply copy-paste augmentations to a set of augmented images so as to synthesise a complex image containing multiple objects. We set the maximum number of possible objects in an image as 10, which means that we randomly pick from 1 to 10 images and copy-paste an object region of each image represented by its pseudo-mask (obtained by SelfMask [9]).

Algorithm 1 Pseudo-code for ZUTIS (using PyTorch-like syntax)

Input. a CLIP image encoder $\psi_{\mathcal{I}}^{enc}$, a transformer decoder $\psi_{\mathcal{I}}^{dec}$, a CLIP text encoder $\psi_{\mathcal{T}}$, two feed-forward networks `FFN`, a projection matrix $\mathbf{W} \in \mathbb{R}^{e_t \times e_v}$, an image $x \in \mathbb{R}^{3 \times H \times W}$, a set of concepts $\mathcal{C}$, queries $Q \in \mathbb{R}^{n_q \times e_v}$, threshold t, temperature τ

Output. predictions for semantic segmentation and instance segmentation

```
# extract dense image features
img_feat = ψ_I^enc(x) # hxwxe_v

# extract text features
text_emb = l2_normalize(ψ_T(C), dim=1) # |C|xe_t

# mask proposals
V = FFN(img_feat.detach()) # hxwxe_v
Q = l2_normalize(FFN(ψ_I^dec(Q, V)), dim=1) # n_qxe_v
M = sigmoid(mm(Q, V.permute(2, 0, 1))) # n_qxhxw

# inference for semantic segmentation
# project image features into the text space
semantic_img_feat = layer_norm(mm(W, img_feat.permute(2, 0, 1))) # e_txhxw
semantic_img_feat = l2_normalize(semantic_img_feat, dim=0)
semantic_prediction = argmax(mm(text_emb, semantic_img_feat), dim=0) # hxw

# inference for instance segmentation
B = M > t # binary masks, n_qxhxw
mask_sizes = sum(B, dim=(1, 2)) # n_q
obj_scores = sum(B * M, dim=(1, 2))/mask_sizes # objectness scores, n_q
avg_feat = sum(
  semantic_img_feat.unsqueeze(dim=0)*B.unsqueeze(dim=1), dim=(2, 3)
)/mask_sizes.unsqueeze(dim=1) # n_qxe_t
avg_feat = l2_normalize(avg_feat, dim=1)
logits = mm(avg_feat, text_emb.t()) # n_qx|C|
conf_scores = max(sigmoid(logits*τ), dim=1)*obj_scores # confidence scores, n_q
mask_classes = argmax(logits, dim=1) # a category label for each mask, n_q
```

mm:matrix multiplication, e_v: a dimension of image features, e_t: a dimension of text embeddings, n_q: the number of queries

C. Additional ablation studies

In this section, we conduct further ablation studies about the influence of using copy-paste augmentation and a hyperparameter choice for temperature, which is used to compute a confidence score of a mask proposal. As in the main paper, we use the VOC2012 [3] `tranval` split for the ablation studies.

C.1. Effect of copy-paste

As mentioned in Sec. 4.1, we apply copy-paste augmentation to synthesise an image with multiple objects following [11]. In Tab. 1, we show that it improves performance of ZUTIS by a large margin with regards to both mIoU (semantic segmentation) and $\text{AP}_{50}^{\text{mk}}$ (instance segmentation).

C.2. Effect of temperature

We compute a confidence score of the mask as a multiplication between the average value of the mask regions and the maximum class probability (see Algorithm 1). For the latter, we consider a temperature parameter τ multiplied to logits for the following sigmoid. In Tab. 2, we evaluate our model with different values for τ and observe that setting τ as 5 yields the

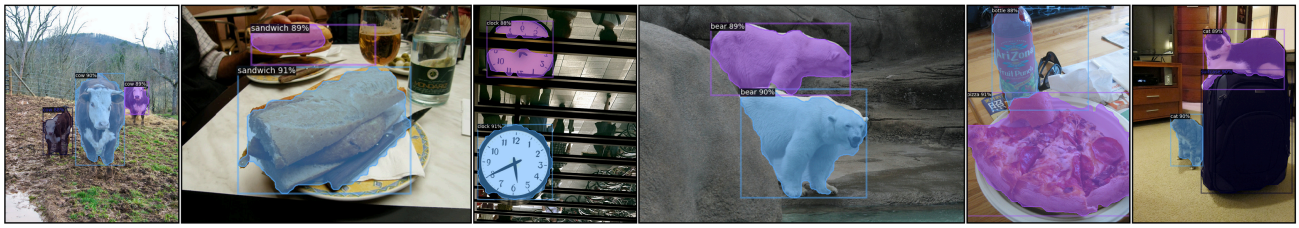

Figure 1. Successful cases of ZUTIS for instance segmentation on COCO-20K. Confident predictions are shown. Zoom in for details.

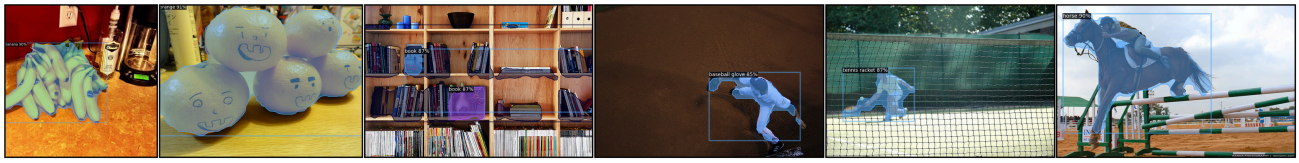

Figure 2. Typical failure cases of ZUTIS on COCO-20K. (Left half) The model fails to distinguish instances of a same category. (Right half) It struggles to differentiate a category from another which is likely to appear together.

copy-paste	mIoU	AP^{mk}_{50}
✗	56.1	28.2
✓	**63.7** (+7.6)	**30.9** (+2.7)

Table 1. Using copy-paste augmentation allows better performance in terms of both mIoU (for semantic segmentation) and AP^{mk}_{50} (for instance segmentaton) on VOC2012 `trainval`.

best performance. For this reason, we use $\tau = 5$ throughout our experiments in the main paper.

τ	0.1	0.5	1	5	10
AP^{mk}_{50}	29.4	30.3	30.8	**30.9**	30.1

Table 2. Effect of temperature τ on instance segmentation performance of ZUTIS.

D. Additional visualisations

We visualise additional instance segmentation results of ZUTIS for successful cases in Fig. 1 and common failure cases in Fig. 2. We note that our model tends to fail in two situations when (i) an image retrieved for a concept is likely to contain multiple instances for the concept (*e.g.*, bananas), or (ii) a given concept is inclined to appear with another category (*e.g.*, a "baseball glove" with a "person" wearing it). We conjecture that this is caused by a lack of high purity images (*i.e.* ones that only contain an object of a single category) for each concept in an image index dataset and/or a use of prompt engineering which is not geared towards curating high purity images from the index dataset.

References

[1] 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

[2] 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. 1

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 2

[4] 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. 1

[5] 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. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[7] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021. 1

[8] 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

[9] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *CVPRW*, 2022. 1

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

[11] Gyungin Shin, Weidi Xie, and Samuel Albanie. Namedmask: Distilling segmenters from complementary foundation models. In *CVPRW*, 2023. 1, 2