

Co-Segmentation without any Pixel-level Supervision with Application to Large-Scale Sketch Classification

– Supplementary Material –

Nikolaos-Antonios Ypsilantis [✉\[0000–0002–3322–6925\]](mailto:ny@fel.cvut.cz)
and Ondřej Chum [✉\[0000–0001–7042–1810\]](mailto:ochum@fel.cvut.cz)

Visual Recognition Group, FEE, Czech Technical University in Prague
[✉ ypsilnik@fel.cvut.cz](mailto:ypsilnik@fel.cvut.cz)

In the Supplementary material, we include an Appendix with qualitative and quantitative results of the proposed co-segmentation method on samples from classes that are not part of ImageNet, as well as the implementation details of the proposed method.

A Appendix

A.1 Out-of-ImageNet classes

We include additional examples of co-salient object detection on the CoCA dataset from classes that are not part of the ImageNet dataset. That is to show qualitatively how the proposed method works on classes outside of the pretraining data of the backbones it utilizes. The examples are shown in Figs. 1 and 2.

We perform additional experiments to support the claim quantitatively. First, we split the CoCA dataset into ImageNet and non-ImageNet classes (21 and 59, respectively). This was achieved by comparing the class labels with BERT [3] embeddings and manually verifying the matches. The results evaluated on the non-ImageNet part of the dataset, comparing the proposed method and the baselines, are shown in Table 1.

A generalization to unseen objects can also be interpreted as how well the co-segmentation works on images that are not used to build the model. To this end, we performed an additional leave-one-out experiment. Each image in the CoCA dataset is segmented by a class-relevance model that is obtained from all other class examples (excluding the tested image). The overall performance is not affected: 0.124 MAE , 0.531 F_{β}^{max} , 0.670 S_{α} . This experiment demonstrates generalization beyond training images.

A.2 Implementation details and timings

For all co-segmentations, images are resized to 256×256 pixels. We use DINO ViT-Small/8 [1] and ImageNet ViT-Small/16 [2] (the latter is not publicly available for the patch size of 8). For the Im4Sketch dataset of the sketch classification application, we use 90 images per class to calculate ξ in equation (1) of the main

Table 1.

Method	non-ImageNet CoCA		
	$MAE \downarrow$	$F_{\beta}^{\max} \uparrow$	$S_{\alpha} \uparrow$
Amir et al. [2]	0.253	0.391	0.550
N-cut [46]	0.145	0.478	0.635
CBNC (ours)	0.128	0.525	0.666

paper, as we have observed that using more shows no advantage. We set the hyperparameter $\tau = 0.2$ similar to [4], $\gamma = 10^{-4}$, $\epsilon = 10^{-5}$ and use 16 eigenvectors to form the biased N-cut vector. The value of the temperature β in the softmax is set to 0.5.

The time needed to extract the ViT features for a set of 90 images is approximately 4 seconds on an NVIDIA GTX TITAN X GPU. Calculating the first eigenvector on ViT features from 90 images for class relevance takes around 0.7 seconds on an Intel Xeon E5-2620 v3 CPU. The estimation of the mask by the biased N-cut takes approximately 0.8 seconds for each image on the same CPU.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT (2019)
4. Wang, Y., Shen, X., Hu, S.X., Yuan, Y., Crowley, J.L., Vaufreydaz, D.: Self-supervised transformers for unsupervised object discovery using normalized cut. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)






Class Name	Sample RGB	Predicted mask
chopsticks		
macaroon		
clover		
persimmon		
rocking horse		

Fig. 1. Examples of co-segmentation for classes that are not part of the ImageNet dataset, which coincides with the pretraining dataset of the backbones utilized by our method.





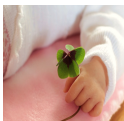



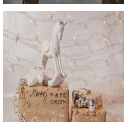



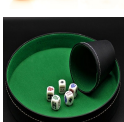
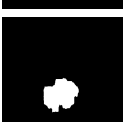
Class Name	Sample RGB	Predicted mask
chopsticks		
pinecone		
clover		
persimmon		
rocking horse		
high heels		
dice		

Fig. 2. Examples of co-segmentation for classes that are not part of the ImageNet dataset, which coincide with the pretraining dataset of the backbones utilized by our method.