— Supplementary Material — PEEKABOO: Hiding Parts of an Image for Unsupervised Object Localization

Hasib Zunair
hasibzunair@gmail.com
A. Ben Hamza
hamza@ciise.concordia.ca
Concordia University
Montreal, QC, Canada

1 Implementation Details

Data preprocessing. Images and masks are resized to 224×224 and normalized using mean and standard deviation of ImageNet. Similar to SelfMask [9], we apply basic data augmentation techniques, including random scaling within the range of $[0.1, 3.0]$ and Gaussian blurring with a probability of 0.5. The parameters of the bilateral solver are the same as those provided in [2].

Architecture. We construct PEEKABOO by using a frozen ViT-S/8 [4] architecture pretrained using DINO [3] as an encoder for feature extraction, from the last attention layer, with a lightweight segmenter head that is a single 1×1 convolutional layer decoder having only 770 learnable parameters.

Model Training. PEEKABOO is trained in a single stage, requiring only a collection of images. We use the Adam optimizer with a learning rate schedule to minimize the total loss function. In addition to the total loss, we also compute the binary cross-entropy loss between the raw predicted masks from the unsupervised segmenter branch and their binarized version. This encourages the predicted soft masks to closely resemble their binarized counterparts. We set the trade-off hyperparameter to 4. The model is trained for 500 iterations on only $10,553$ images from the DUTS-TR dataset [12] with a batch size of 50, which corresponds to slightly more than 2 epochs. For image masking, we utilize the Irregular Mask (IMs) dataset [5], containing masks with random streaks and holes of various shapes, as illustrated in Figure 1.

Model Testing. After training, given an input image, the model simply makes a prediction by outputting a segmentation mask for the salient object(s). During inference, the input image undergoes normalization only. Moreover, there is no random masking procedure applied, as is done during the training stage.

Hardware and software details. The experiments were performed on a Linux workstation running 4.8Hz and 64GB RAM, equipped with a single NVIDIA RTX 3080Ti GPU featuring 12GB of memory. All algorithms are implemented using the PyTorch framework.

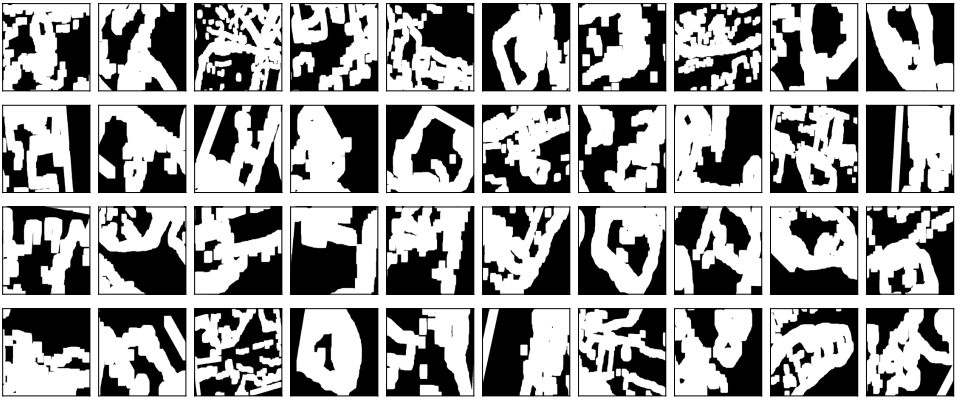

Figure 1: **Visualization of masks during training in PEEKABOO**. Some masks cover more than 50% of the image. Images are from Irregular Masks Dataset [5] after applying binary thresholding.

2 Additional Visual Comparison Results

In Figure 2, we present additional experimental results on unsupervised object localization to further demonstrate the effectiveness of PEEKABOO in localizing salient objects. As can be see, PEEKABOO consistently excels in localizing salient objects across all datasets. Its performance is particularly noteworthy when dealing with small objects, reflective surfaces, and objects situated against complex or dimly illuminated backgrounds. This capability highlights PEEKABOO's robustness and adaptability in various challenging scenarios, further validating its superiority over strong baselines in unsupervised object localization tasks.

3 Method Cost Discussion

We compare PEEKABOO against training-free and training-based methods, which we find have significantly different costs at both training and inference time. Specifically, we demonstrate the efficiency of our method. PEEKABOO is a segmenter head, on top of a frozen Self-distillation with No Labels (DINO) [3] as an encoder, which consists of a lightweight **single** 1×1 **convolutional layer** decoder having only **770 learnable parameters**. The model is trained for **2 epochs** on $10,553$ **images** from the DUTS-TR dataset [12] on a **single consumer grade NVIDIA RTX 3080Ti GPU**.

Inference with training-free methods like TokenCut [14] and Deep Spectral Methods (DSMs)[7] is slowed down due to the expensive computation of the Laplacian matrix eigenvectors. LOST[10], while somewhat faster, notably lags behind PEEKABOO in terms of localization performance.

Among training-based methods, FreeSOLO [13] stands out with approximately 66 million learnable parameters, trained over 241 thousand unlabeled images for 60 thousand iterations across 8 GPUs, making its training considerably more resource-intensive compared to ours. Also, its backbone is pretrained on ImageNet with 1.28 million unlabeled images. SelfMask [9] utilizes 36 million learnable parameters and trains for 12 epochs on the DUTS-TR dataset [12]. COMUS [15] requires three days of training on two 8-GPU servers for

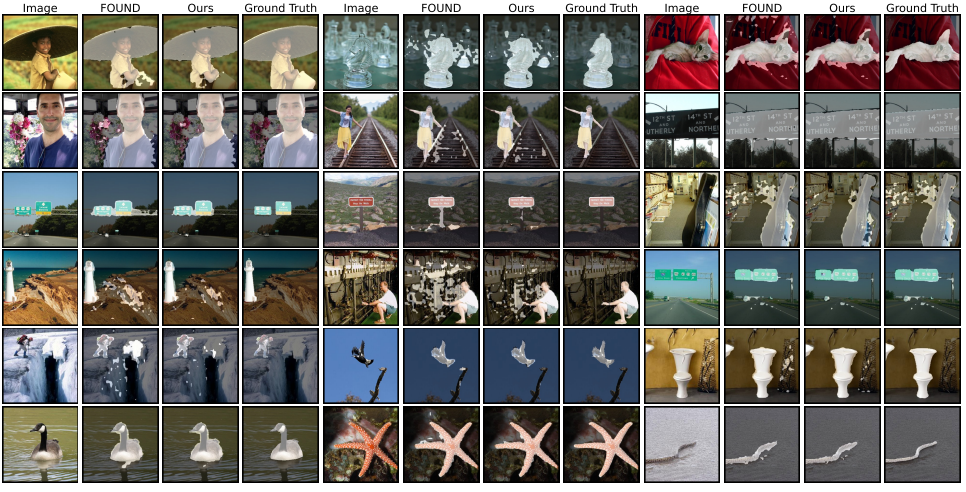

Figure 2: **More examples of visual comparison of PEEKABOO and state-of-the-art FOUND [11] on ECSSD, DUT-OMRON and DUTS-TE datasets**. Across all datasets, PEEKABOO excels in localizing salient objects, particularly when they are small, reflective, or situated against complex or dimly illuminated backgrounds. Zoom in to observe the results more closely.

its heavy segmentation backbone. DINOSAUR [8] is trained on over 300 thousand images from synthetic and real-world sources and demands 8 GPUs for training. DeepCut [1] has 30K learnable parameters, and WSCUOD [6], which incorporates a DINO-ViT-S/16 backbone, consists of 2 million learnable parameters and requires training on 6 GPUs, making it substantially more expensive to train compared to our method.

References

[1] Amit Aflalo, Shai Bagon, Tamar Kashti, and Yonina Eldar. DeepCut: Unsupervised segmentation using graph neural networks clustering. In *Proc. IEEE International Conference on Computer Vision*, pages 32–41, 2023.

[2] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *Proc. European Conference on Computer Vision*, pages 617–632, 2016.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. IEEE International conference on Computer Vision*, pages 9650–9660, 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[5] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proc. European Conference on Computer Vision*, pages 85–100, 2018.

[6] Yunqiu Lv, Jing Zhang, Nick Barnes, and Yuchao Dai. Weakly-supervised contrastive learning for unsupervised object discovery. *arXiv preprint arXiv:2307.03376*, 2023.

[7] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022.

[8] Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *International Conference on Learning Representations*, 2023.

[9] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2022.

[10] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *Proc. British Machine Vision Conference*, 2021.

[11] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobeckỳ, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3176–3186, 2023.

[12] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017.

[13] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. FreeSOLO: Learning to segment objects without annotations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 14176–14186, 2022.

[14] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022.

[15] Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *International Conference on Learning Representations*, 2023.