

PEEKABOO: Hiding Parts of an Image for Unsupervised Object Localization

Hasib Zunair
hasibzunair@gmail.com

Concordia University
Montreal, QC, Canada

A. Ben Hamza
hamza@ciise.concordia.ca

Abstract

Localizing objects in an unsupervised manner poses significant challenges due to the absence of key visual information such as the appearance, type and number of objects, as well as the lack of labeled object classes typically available in supervised settings. While recent approaches to unsupervised object localization have demonstrated significant progress by leveraging self-supervised visual representations, they often require computationally intensive training processes, resulting in high resource demands in terms of computation, learnable parameters, and data. They also lack explicit modeling of visual context, potentially limiting their accuracy in object localization. To tackle these challenges, we propose a single-stage learning framework, dubbed PEEKABOO, for unsupervised object localization by learning context-based representations at both the pixel- and shape-level of the localized objects through image masking. The key idea is to selectively hide parts of an image and leverage the remaining image information to infer the location of objects without explicit supervision. The experimental results, both quantitative and qualitative, across various benchmark datasets, demonstrate the simplicity, effectiveness and competitive performance of our approach compared to state-of-the-art methods in both single object discovery and unsupervised salient object detection tasks. Code and pre-trained models are available at: <https://github.com/hasibzunair/peekaboo>

1 Introduction

Localizing objects in images or videos is a fundamental task in many real-world computer vision systems, including autonomous driving and robotics. The goal is to detect/segment the most salient objects in an image. High-performing methods for object localization typically rely on large cohorts of human-annotated datasets for supervised learning [19]. However, this learning paradigm faces two main significant limitations. First, acquiring annotated datasets is notoriously time-consuming and costly, as well as prone to errors due in part to annotators' fatigue. Second, the finite and predefined nature of object classes limits the scope of what supervised models can localize, rendering them ineffective for segmenting novel objects.

In recent years, several approaches have been proposed to tackle the daunting task of unsupervised object localization, which is inherently challenging. This difficulty arises from the need to identify salient objects without prior knowledge of their appearance or reliance on human-annotated datasets. Some approaches for unsupervised object discovery leverage

region proposal techniques and representer point selection [32, 64, 48], while others rely on adversarial training strategies [23, 87]. Moreover, some methods refine mask proposals [39, 44], albeit at the expense of requiring millions of learnable parameters, extensive unlabeled image datasets, and multi-stage training procedures. Drawing upon the localization properties of self-supervised visual representations [0], training-free graph-based methods [24, 60, 40] have demonstrated robust performance across various tasks and benchmarks. However, these approaches often struggle in practical real-world scenarios due largely to variations in illumination, reflections on water and shiny surfaces [61], overlapping objects [30], occlusions of small objects, complex backgrounds, assumptions that the largest object is salient, and segmenting multiple objects as one [22, 24, 40]. Furthermore, they may erroneously count multiple instances of the same object as one, leading to over- or under-segmentation, as well as segment non-salient regions and produce noisy and discontinuous predictions. In addition, training-based methods impose significant computational demands [29], rely on extensive unlabeled image datasets [27], employ combinations of multiple networks, ensemble methods, and require test-time training [0, 29]. Overcoming these challenges is essential for deploying reliable vision systems in real-world scenarios, such as autonomous driving, where accuracy and efficiency are paramount. Therefore, there is a pressing need to develop unsupervised object localization methods that are both accurate and computationally efficient.

To address the aforementioned challenges, we introduce PEEKABOO¹, a novel single-stage learning framework for unsupervised object localization by adopting a context-based representation learning strategy to enhance foreground-background discrimination, without resorting to pixel-level reconstruction. Given an image and its heavily masked counterpart, our model first predicts pixel-level semantic labels for both the unmasked and masked inputs. This process exploits contextual information by utilizing nearby non-masked pixels to inform predictions for masked pixels, followed by learning shape-level context by ensuring consistency between the object mask predictions for the masked and unmasked inputs. The main contributions of this work can be summarized as follows: (1) We propose a learning paradigm that aims to explicitly model context through hiding parts of an image for unsupervised object localization; (2) We show through experimental results and ablation studies on six benchmark datasets that PEEKABOO yields competitive performance against state-of-the-art methods in both single object discovery and unsupervised saliency detection tasks.

2 Related work

Unsupervised Object Localization. Generative representation learning methods like the one proposed by Melas-Kyriazi *et al.* [23] employ a two-stage adversarial training approach, utilizing BigBiGAN [0] to create a synthetic dataset, followed by segmenter training. Another line of work involves generating mask proposals that are refined through training. FreeSOLO [39] proposes Free Mask to generate coarse object masks refined via self-training, but requires several millions of learnable parameters and large-scale unlabeled image datasets. More recent efforts leverage the strong localization properties of DINO [0], a self-supervised pre-training method, with or without additional training. For instance, LOST [30] is training-free and constructs a weighted graph of patches, segmenting foreground objects based on their similarity, but struggles in scenarios involving object overlap or when objects cover

¹The name PEEKABOO is inspired by the children’s game of hiding one’s face and then revealing it suddenly.

most of the image. TokenCut [41] and Deep Spectral Methods (DSMs) [24] are graph-based methods, but their performance is hindered by the computational overhead of computing the eigenvectors of the Laplacian matrix. They often assume a single salient object occupies the foreground and may count multiple instances of the same object as one. Moreover, they struggle with small and occluded objects, leading to over- or under-segmentation. Self-Mask [29] is training-based method, which employs an ensemble of three self-supervised vision encoders (SwAV [6], DINO [7], MoCov2 [40]) and spectral clustering to generate more than twenty salient masks for each image, making the training considerably more expensive. DINOSAUR [27] learns representations from hundreds of thousands of unlabeled images by separating the features of an image and reconstructing them into individual objects or parts. WSCUOD [22] proposes a semantic-guided method that extracts object-centric representations to generate the final object regions from the image. DeepCut [8] utilizes features extracted by DINO [7] and trains a graph neural network to generate segmentation masks, incorporating test-time training on each image. In contrast to these approaches, our method avoids heavy and complex training procedures such as large-scale adversarial training, multi-stage training, learning millions of parameters, and test-time training.

Masked Image Modeling for Vision. While early approaches for masked image modeling focused on reconstructing missing parts from masked images [21, 26], recent methods such as Masked Autoencoders [16] and DINO [7] have aimed to predict visual tokens in a self-supervised manner [2]. Also, masking has recently been utilized as supervisory signals for models to learn context-based representations in fully supervised learning settings, facilitating accurate perception of 2D [49] and 3D [20] visual data. However, the application of masked modeling tailored for unsupervised learning, specifically for discovering novel objects from unlabeled data, remains relatively unexplored. Moreover, our approach distinguishes itself from existing methods by making pixel-level predictions instead of predicting missing parts of the input itself (i.e., pixel or visual token).

3 Method

We begin by formulating the task at hand and subsequently introduce the main components of the proposed masked unsupervised object localization framework, as illustrated in Figure 1.

Problem Statement. We address two main tasks in unsupervised object localization from images: single object discovery and unsupervised saliency detection. The former involves enclosing the main object or one of the main objects of interest within a bounding box, while the latter aims to generate a binary mask where foreground objects are assigned a value of 1 and the background is assigned 0, thereby highlighting the object(s) of interest in the image. Both tasks are class-agnostic, i.e., they are not restricted to a predefined set of object classes.

3.1 Masked Unsupervised Object Localization

Self-Training Unsupervised Segmenter. We design an unsupervised segmenter model f_θ by leveraging the frozen Self-distillation with No Labels (DINO) [7] as an encoder for feature extraction with a lightweight single 1×1 convolutional layer decoder having only 770 learnable parameters. The model is trained end-to-end in a self-supervised fashion, and learns to predict a refined version of its own predictions from the input image \mathbf{I} . The predicted mask \mathbf{M}_p is refined by applying image binarization, followed by bilateral solver [2], an edge-aware smoothing technique that enhances mask quality. The refined mask, denoted

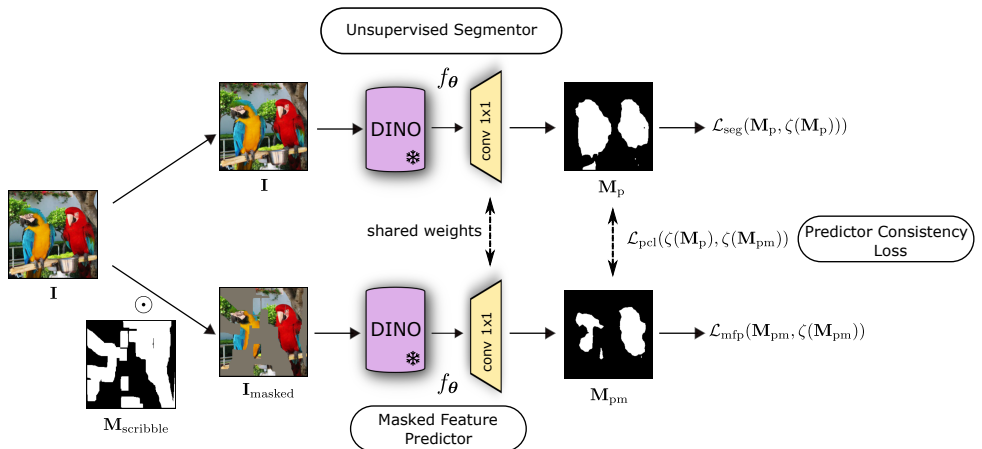


Figure 1: Overview of PEEKABOO framework for unsupervised object localization. The proposed learning paradigm consists of an Unsupervised Segmentor, a Masked Feature Predictor and a Predictor Consistency Loss. Here, f_θ is a frozen Self-distillation with No Labels (DINO) encoder [14] paired with a lightweight trainable 1×1 convolutional layer decoder having only 770 learnable parameters. The two branches are identical and share weights. After training, f_θ is utilized to generate class-agnostic predicted segmentation masks.

as $\zeta(\mathbf{M}_p)$, serves as pseudo ground-truth for self-training, with $\zeta(\cdot)$ representing the image binarization and bilateral solver operation. We train f_θ by minimizing the segmenter loss $\mathcal{L}_{\text{seg}}(\mathbf{M}_p, \zeta(\mathbf{M}_p))$, which is the binary cross-entropy between the predicted mask and its refined version. Application-specific loss functions can also be used in lieu of cross-entropy to further improve performance depending on the use-case.

Masked Input. Incorporating image masking into context-based representation learning is intuitive as the masked image retains some contextual information, requiring the model to capture background knowledge about the image. To facilitate this, we leverage the Irregular Mask (IMs) dataset [21], which is a collection of masks with random streaks and holes of arbitrary shapes commonly used for image inpainting [21, 63]. From this dataset, we create a subset of images containing only large masks by filtering the masks based on the percentage p of zero pixels they contain. Masks with p greater than 50% are included, while others are discarded. In our experiments, we observe that these large masks generally enhance localization performance compared to smaller ones. During PEEKABOO training, we randomly sample a large mask, perform binary thresholding to set pixel values to either 0 or 1, and denote this mask as $\mathbf{M}_{\text{scribble}}$. We generate a masked image $\mathbf{I}_{\text{masked}} = \mathbf{I} \odot \mathbf{M}_{\text{scribble}}$, where \mathbf{I} is the original input image and \odot denotes element-wise multiplication. Essentially, the masked image retains a layout akin to the original input, albeit with substantial portions of the pixels removed.

Masked Feature Predictor (MFP). We introduce MFP to explicitly learn context-based representations. MFP is tasked with generating pixel-level semantic labels, distinguishing between foreground and background, for the masked portions of the image. This prompts the model to learn contextual information by utilizing the available pixel-level details of neighboring non-masked pixels to predict the label for each masked pixel. Given the masked image $\mathbf{I}_{\text{masked}}$, we train f_θ to produce the predicted mask \mathbf{M}_{pm} . It is important to note that

f_θ is a Siamese network-like architecture [6], where the branches are identical and share weights, and is trained by minimizing the binary cross-entropy loss $\mathcal{L}_{\text{mfp}}(\mathbf{M}_{\text{pm}}, \zeta(\mathbf{M}_{\text{pm}}))$.

Predictor Consistency Loss (PCL). The objective of PCL is to ensure consistency between predictions for the input image and its masked version. By aligning these predictions, PCL learns shape-level representations that complement the pixel-level representations learned by MFP. This alignment is achieved by ensuring that the predictions of the masked input match those of the unmasked input. Since both predictions are generated by the same model f_θ , this fosters invariance to images with partial inputs, ultimately leading to improved localization performance. More specifically, we maximize the similarity between the predictions from the unsupervised segmenter and masked feature predictor by minimizing the L_2 -loss $\mathcal{L}_{\text{pcl}} = \|\zeta(\mathbf{M}_{\text{p}}) - \zeta(\mathbf{M}_{\text{pm}})\|^2$.

Overall Loss Function. Using the unsupervised segmenter, masked feature predictor and predictor consistency loss, we define the overall loss function as follows:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{seg}}(\mathbf{M}_{\text{p}}, \zeta(\mathbf{M}_{\text{p}})) + \mathcal{L}_{\text{mfp}}(\mathbf{M}_{\text{pm}}, \zeta(\mathbf{M}_{\text{pm}})) + \mathcal{L}_{\text{pcl}}(\zeta(\mathbf{M}_{\text{p}}), \zeta(\mathbf{M}_{\text{pm}})), \quad (1)$$

where the scalar α is a nonnegative trade-off hyperparameter, which controls the contribution of the segmenter loss and is set to 3/2. During training, $\mathcal{L}_{\text{total}}$ is minimized between predictions and pseudo ground-truth labels using stochastic gradient descent for several epochs to learn the parameters of f_θ . Similar to [29, 51], we use only 10,553 images from the DUTS-TR dataset [58] for training PEEKABOO. For inference, given a test image \mathbf{I} , the trained network f_θ is employed in unsupervised object localization to obtain segmentation masks.

4 Experiments

In this section, we evaluate the performance of PEEKABOO in comparison with state-of-the-art methods. Details on the implementation, architecture and training, as well as method cost discussion and additional experimental results are included in the supplementary material.

4.1 Experimental Setup

Datasets. For single object discovery, we conduct experiments on PASCAL VOC07 [12] and VOC12 [13], which consist images of single large object, and also on COCO20K [19], which includes two and often dozens of objects. For the unsupervised saliency detection task, we evaluate the performance of our approach on three popular saliency detection benchmark datasets: DUT-OMRON [43], DUTS-TE [58] and ECSSD [28]. These datasets encompass a diverse array of objects set against various backgrounds.

Baselines. We evaluate the performance of our method against several state-of-the-art training-free methods, including LOST [30], TokenCut [40] and Deep Spectral Methods [24]. We also compare against training-based generative methods like BigBiGAN [57] and Melas-Kyriazi et al. [23], along with Transformer-based methods such as FOUND [51], Self-Mask [29], FreeSOLO [59], and DINOSAUR [27]. Furthermore, we compare against more recent methods such as WSCUOD [22] and DeepCut [1]. Notably, these recent approaches typically entail multiple stages of training, learning several millions of parameters, test-time training, combining multiple learnable networks, model ensembling, and require training on hundreds of thousands of images.

Evaluation Metrics. For single object discovery, we report results using the Correct Localization (CorLoc) metric [60, 83], which measures the percentage of correct boxes, where predicted boxes have an intersection-over-union (IoU) greater than 0.5 with one of the ground-truth boxes. For unsupervised saliency detection, we report results in terms of IoU, pixel wise accuracy (Acc) and maximal F_β score ($\max F_\beta$).

4.2 Results and Analysis

Single Object Discovery Results. We evaluate the performance of PEEKABOO against several state-of-the-art methods on three single object discovery datasets, and the results are summarized in Table 1. We use PEEKABOO with a ViT-S/8 [42] architecture pre-trained with DINO [2]. As reported in Table 1, PEEKABOO demonstrates superior performance compared to all training-free methods, including LOST [60], TokenCut [40], and DSM [24]. Moreover, PEEKABOO maintains computational efficiency by avoiding the computation of eigenvectors, thereby ensuring faster inference [24, 40]. This demonstrates the efficiency of PEEKABOO for real-time object discovery. PEEKABOO also performs better than training-based methods (SelfMask [29], FreeSOLO [39], FOUND [81], WSCUOD [22] and DeepCut [1]), while being notably simpler to train. In addition, it outperforms DINOSAUR [27] on VOC07 and VOC12, while DINOSAUR [27] performs better on COCO20K. However, it is worth noting that DINOSAUR [27] incurs an extensive training cost, as it employs a significantly larger ViT-based architecture trained on over three hundred thousand images sourced from both synthetic and real-world datasets. Evidently, PEEKABOO outperforms DINOSAUR [27] on VOC12, achieving a relative improvement of 7.8% in terms of Cor-Loc. It also performs better than DeepCut [1], which employs multiple learnable networks (ViTs and GNNs) and involves test-time training. PEEKABOO stands out for its accuracy and efficiency, boasting approximately 42,857 times fewer learnable parameters compared to SelfMask [29], which relies on an ensemble of three self-supervised vision encoders. Similarly, it is significantly more parameter-efficient than FreeSOLO [39], which employs over two hundred thousand images and 66 million learnable parameters, roughly 24 and 85,714 times higher than our method, respectively. These results underscore the accuracy and efficiency of our method in single object discovery.

Unsupervised Saliency Detection Results. In Table 2, we present the performance comparison across three popular saliency detection benchmarks. PEEKABOO demonstrates superiority over training-based techniques such as BigBiGAN [57] and Melas-Kyriazi et al.[23], which rely on large-scale adversarial training. Notably, PEEKABOO consistently outperforms training-free methods, including LOST[60], TokenCut [40], and DSM [24] across all evaluation metrics. Interestingly, PEEKABOO achieves these results with computational efficiency, even when compared to approaches incorporating bilateral solver post-processing [9]. Moreover, it demonstrates superior performance compared to SelfMask [29], except for IoU on DUT-OMRON, despite the substantial training cost associated with Self-Mask. This cost stems from its ensemble approach, utilizing three self-supervised vision encoders (SwAV [6], DINO [2], MoCov2 [10]), and generating twenty masks for each image during training. Furthermore, PEEKABOO outperforms DeepCut [1], which employs multiple learnable networks (ViTs and GNNs) and requires test-time training. It also consistently outperforms WSCUOD [22] across all datasets and evaluation metrics, even when WSCUOD employs bilateral solver post-processing. Notably, WSCUOD utilizes weak supervisory signals for training, whereas PEEKABOO operates directly on unlabeled images

Table 1: Performance comparison of our method and state-of-the-art methods in single object discovery using CorLoc as evaluation metric. The best results are in bold, while the second-best are underlined. Higher values indicate better results. “+CAD” denotes an additional second-stage class-agnostic detector trained with unsupervised pseudo-box labels.

Method	Learning	VOC07	VOC12	COCO20K
Zhang et al. [46]		46.2	50.5	34.8
DDT+ [41]		50.2	53.1	38.2
rOSD [35]		54.5	55.3	48.5
LOD [36]		53.6	55.1	48.5
DINO [7]		45.8	46.2	42.1
LOST [30] (ViT-S/16)		61.9	64.0	50.7
LOST + CAD [30]		65.7	70.4	57.5
DSM [24] (ViT-S/16)		62.7	66.4	52.2
TokenCut [40] (ViT-S/16)		68.8	72.1	58.8
TokenCut + CAD [40]		71.4	75.3	62.6
SelfMask [29]	✓	<u>72.3</u>	75.3	62.7
FOUND† [31]	✓	71.7	<u>75.6</u>	61.1
FreeSOLO [39]	✓	56.1	56.7	52.8
DeepCut [11]	✓	69.8	72.2	61.6
WSCUOD [22]	✓	70.6	72.1	63.5
DINOSAUR [27]	✓	-	70.4	67.2
PEEKABOO (ViT-S/8) (Ours)	✓	72.7	75.9	<u>64.0</u>

and has significantly fewer learnable parameters, approximately 2597 times less. In addition, even with bilateral solver post-processing, PEEKABOO maintains its superior performance, demonstrating its ability to produce high-quality object masks without heavy reliance on post-processing techniques for enhanced results.

Qualitative Results. In Figure 2, we present visual examples showcasing the predictions generated by our PEEKABOO model and the FOUND [31] baseline. The first two rows, featuring samples from the ECSSD dataset, illustrate that the baseline tends to over-segment salient objects and struggles with cases involving reflections on water and shiny surfaces, resulting in nearly duplicated segmentation. Moving to the middle two rows, which depict samples from the DUT-OMRON dataset, we observe the baseline’s challenges with complex backgrounds, dark scenes, and small objects, along with its tendency to segment non-salient regions and produce noisy predictions. Finally, the last two rows, representing instances from the DUTS-TE dataset, highlight the baseline’s difficulty in localizing small objects and its tendency to generate noisy and discontinuous predictions. By comparison, PEEKABOO excels in localizing salient objects under various conditions, including when they are small, reflecting on water or shiny surfaces, or in complex or low-illumination backgrounds. Moreover, PEEKABOO avoids common pitfalls such as over-segmentation of salient objects, segmentation of non-salient regions, and generation of noisy predictions. This is because MFP operates on partial inputs through masking, thereby generating context-based pixel-level representations. PCL, on the other hand, complements MFP by learning shape-level representations through aligning predictions from both unmasked and masked inputs.

Table 2: **Performance comparison of our method and state-of-the-art methods in unsupervised saliency detection.** “+BS” and “+CRF” indicate the utilization of the post-processing techniques [4] and [17], respectively. The symbol † denotes results reproduced using the publicly available codes.

Method	Learning	DUT-OMRON			DUTS-TE			ECSSD		
		Acc	IoU	max F_β	Acc	IoU	max F_β	Acc	IoU	max F_β
HS [17]		84.3	43.3	56.1	82.6	36.9	50.4	84.7	50.8	67.3
wCtr [4]		83.8	41.6	54.1	83.5	39.2	52.2	86.2	51.7	68.4
WSC [23]		86.5	38.7	52.3	86.2	38.4	52.8	85.2	49.8	68.3
DeepUSPS [25]		77.9	30.5	41.4	77.3	30.5	42.5	79.5	44.0	58.4
BigBiGAN [6]		85.6	45.3	54.9	87.8	49.8	60.8	89.9	67.2	78.2
E-BigBiGAN[6]		86.0	46.4	56.3	88.2	51.1	62.4	90.6	68.4	79.7
Melas-Kyriazi et al. [23]		88.3	50.9	-	89.3	52.8	-	91.5	71.3	-
LOST [6]		79.7	41.0	47.3	87.1	51.8	61.1	89.5	65.4	75.8
DSM [24]		80.8	42.8	55.3	84.1	47.1	62.1	86.4	64.5	78.5
TokenCut [10]		88.0	53.3	60.0	90.3	57.6	67.2	91.8	71.2	80.3
SelfMask [24]	✓	90.1	58.2	-	92.3	62.6	-	94.4	78.1	-
FOUND† [5]	✓	<u>90.7</u>	57.1	<u>79.9</u>	<u>93.5</u>	<u>63.7</u>	<u>85.2</u>	94.9	80.6	<u>95.1</u>
DeepCut [1]	✓	-	-	-	-	59.5	-	-	74.6	-
W SCUOD [24]	✓	89.7	53.6	64.4	91.7	59.9	73.1	92.2	72.7	85.4
PEEKABOO (Ours)	✓	91.5	<u>57.5</u>	80.4	93.9	64.3	86.0	<u>94.6</u>	<u>79.8</u>	95.3
LOST + BS [6]	✓	81.8	48.9	57.8	88.7	57.2	69.7	91.6	72.3	83.7
DSM + CRF [24]	✓	87.1	56.7	64.4	83.8	51.4	56.7	89.1	73.3	80.5
W SCUOD + BS [24]	✓	90.9	58.5	68.3	92.5	63.0	<u>76.4</u>	92.8	74.2	89.6
TokenCut + BS [10]	✓	89.7	<u>61.8</u>	<u>69.7</u>	91.4	62.4	75.5	93.4	77.2	87.4
SelfMask + BS [24]	✓	<u>91.9</u>	65.5	-	93.3	66.0	-	95.5	81.8	-
FOUND + BS† [5]	✓	91.7	60.9	69.1	<u>94.0</u>	<u>66.1</u>	75.0	<u>95.2</u>	<u>81.7</u>	<u>93.0</u>
PEEKABOO + BS (Ours)	✓	92.4	61.2	71.4	94.4	66.3	77.4	94.9	80.6	93.7

4.3 Ablation Study

We analyze how each of the key components of the proposed framework affects the overall performance. We also examine the impact of masking on model performance.

Effectiveness of Masked Feature Predictor (MFP). Figure 3 (left) illustrates the benefit of using MFP, which helps capture local semantics by generating pixel-level predictions for masked portions of the image. This enables the model to learn context by leveraging information from nearby non-masked pixels to make a prediction for the masked pixel. The effectiveness of MFP is evident in the improved localization performance across all datasets. Particularly noteworthy is the significant performance improvement observed on COCO20K, which contains a diverse range of images, including those with multiple objects, compared to the VOC07 and VOC12 datasets, which primarily consist of single large objects.

Effectiveness of Predictor Consistency Loss (PCL). As depicted in Figure 3 (left), PCL contributes to performance improvement across all datasets. By aligning the predictions of the input and its masked counterpart, we can learn representations that are invariant to masking. This enables the model to capture the shapes of objects to be segmented, resulting in more accurate predictions. Notably, the most substantial improvement is observed on COCO20K, underscoring the potential of PEEKABOO in scenarios featuring diverse images

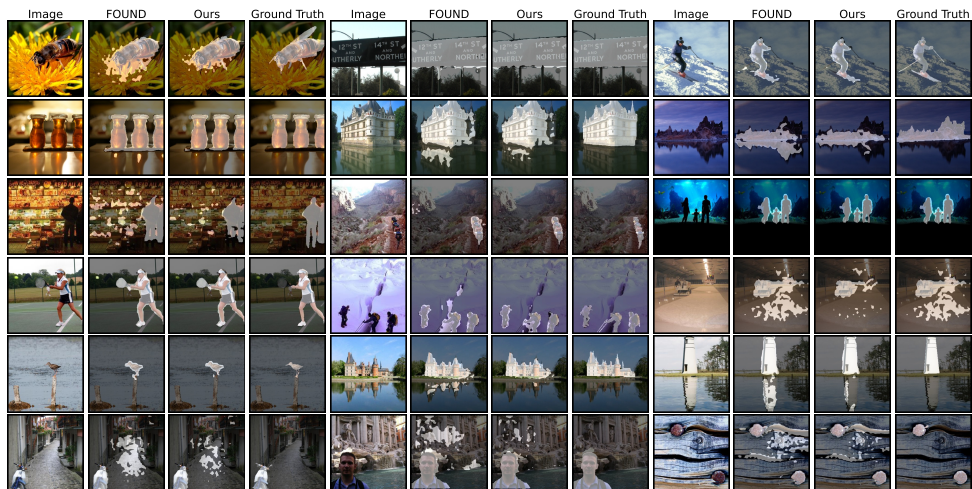


Figure 2: Visual comparison of PEEKABOO and state-of-the-art FOUND [51] on EC-SSD, DUT-OMRON and DUTS-TE datasets. Across all datasets, PEEKABOO excels in localizing salient objects, particularly when they are small, reflective, or situated against complex or dimly illuminated backgrounds. Zoom in to observe the results more closely.

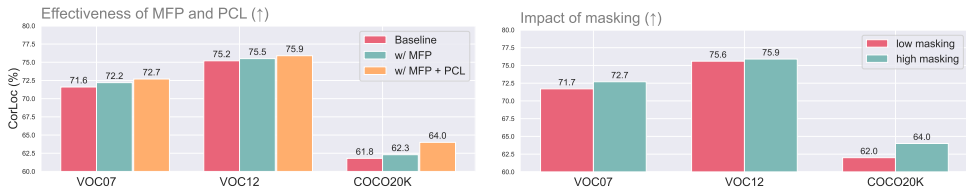


Figure 3: Ablation study of different modules of PEEKABOO (left) and impact of masking (right) using VOC07, VOC12 and COCO20K datasets. MFP and PCL consistently help improve performance. PEEKABOO with high masked pixels yields better performance.

with multiple objects.

Impact of Masking. In Figure 3 (right), we illustrate the impact of masking on model performance during the training of PEEKABOO. Our training strategy involves utilizing subsets of high and low masks, where M_{scribble} in the subset either comprises large or small portions of zero pixels. We observe that localization performance generally improves when a large portion of the image is masked. This observation aligns with intuition since low masking preserves most of the information, leading the model to learn redundant features.

5 Discussion and Limitations

Relation to Self-Supervised Learning via Dual Networks. Self-supervised learning (SSL) serves as a pre-training strategy aimed at mitigating the reliance on labeled data for supervised learning. Typically, SSL involves a pretext task that exploits the inherent semantics and structure of unlabeled data to learn the relationship between input and pseudo-label (i.e., image representations), easily generated from the input itself. Numerous SSL meth-

ods [9, 6, 8, 9, 15, 45] rely on Siamese networks [6] to encourage the similarity between two outputs, showcasing the effectiveness of self-supervised pre-training in various target tasks, including image classification, segmentation, and detection [9].

While this similarity concept resembles the predictor consistency loss introduced in PEEKABOO, there is a key distinction in the type of output: SSL methods yield latent representations (i.e., feature vectors), whereas PEEKABOO outputs prediction labels (i.e., segmentation maps). Moreover, PEEKABOO differs from SSL methods in a number of aspects. First, SSL methods typically involve a pre-training stage on large amounts of unlabeled data followed by fine-tuning for the target task. In contrast, PEEKABOO directly applies to the target task of unsupervised object localization. Second, while SSL methods operate on two randomly augmented samples of the image through rotation, cropping, translation, or blurring [9, 8, 45], PEEKABOO works with the original image and its masked version. Third, unlike methods such as BYOL [15], SwAV [6] and SimSiam [9], which employ two independent networks operating on the two samples, PEEKABOO utilizes a Siamese network with shared weights, where both networks are identical. Fourth, PEEKABOO employs a completely different objective function compared to existing methods: SimCLR [8] uses a contrastive loss requiring both positive and negative pairs, Barlow Twins [45] and VICReg [9] use a cross-correlation matrix based loss, while BYOL [15], SwAV [6] and SimSiam [9] use a cosine similarity loss. In summary, PEEKABOO aims to unify the concepts of self-supervised pretraining [9, 8, 9, 45] on unlabeled data and fine-tuning on a target task into a single framework that models contextual relationship among pixels and learns semantics and structure through image masking. This approach potentially avoids the discrepancies involved in separate pre-training and fine-tuning stages.

Limitations. Figure 4 illustrates some instances where PEEKABOO encounters challenges. One notable limitation is observed in indoor scenes, where PEEKABOO struggles to accurately segment salient objects. Moreover, in cases where salient objects, such as trees, are easily distinguishable visually, PEEKABOO may exhibit difficulties in segmentation, suggesting potential over-generalization issues.

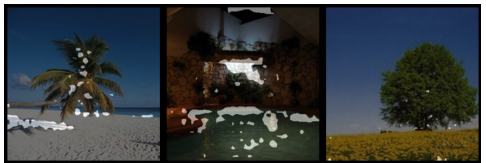


Figure 4: Visualization of failure cases of PEEKABOO on DUT-OMRON for object localization. No refinement step is applied. Zoom in to observe the results more closely.

6 Conclusion

We introduced PEEKABOO, a new single-stage learning paradigm that models contextual relationship among pixels through image masking for unsupervised object localization. The proposed framework learns context-based representations at the pixel-level by making predictions on masked images and at the shape-level by matching the predictions of the masked input to the unmasked one. Through extensive experiments, our quantitative and qualitative results demonstrated that our approach outperforms state-of-the-art methods, excelling in localizing salient objects, especially when they are small, reflect on water or shiny surfaces, or when the background is complex or poorly illuminated. Notably, our method avoids over-segmenting salient objects, segmenting non-salient regions, and producing noisy predictions. For future work, we aim to explore learning context-based representations for direct localization of individual concepts in open-vocabulary semantic segmentation.

Acknowledgments. This work was supported in part by the Discovery Grants Program of the Natural Sciences and Engineering Research Council of Canada under grant RGPIN-2024-04291.

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] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.
- [3] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2021.
- [4] Jonathan T Barron and Ben Poole. The fast bilateral solver. In *Proc. European Conference on Computer Vision*, pages 617–632, 2016.
- [5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, 1993.
- [6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020.
- [7] 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.
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.
- [9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.
- [10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.
- [11] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, 2019.
- [12] 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.

- [13] M Everingham, S.M. Eslami, L Van Gool, CKI Williams, J Winn, and A Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111:98–136, 2015.
- [14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.
- [15] Jean-Bastien Grill, Florian Strub, Florent Alché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.
- [17] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 2011.
- [18] Nianyi Li, Bilin Sun, and Jingyi Yu. A weighted sparse coding framework for saliency detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5216–5223, 2015.
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. European Conference on Computer Vision*, pages 740–755, 2014.
- [20] 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.
- [21] Lizhao Liu, Zhuangwei Zhuang, Shangxin Huang, Xunlong Xiao, Tianhang Xiang, Cen Chen, Jingdong Wang, and Mingkui Tan. CPCM: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation. In *Proc. IEEE International Conference on Computer Vision*, pages 18413–18422, 2023.
- [22] Yunqiu Lv, Jing Zhang, Nick Barnes, and Yuchao Dai. Weakly-supervised contrastive learning for unsupervised object discovery. *arXiv preprint arXiv:2307.03376*, 2023.
- [23] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. In *International Conference on Learning Representations*, 2022.
- [24] 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.

- [25] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems*, 2019.
- [26] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
- [27] 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.
- [28] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended CSSD. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2015.
- [29] 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.
- [30] 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.
- [31] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobecký, 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.
- [32] Yeonghwan Song, Seokwoo Jang, Dina Katabi, and Jeany Son. Unsupervised object localization with representer point selection. In *Proc. IEEE International Conference on Computer Vision*, pages 6534–6544, 2023.
- [33] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022.
- [34] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104: 154–171, 2013.
- [35] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *Proc. European Conference on Computer Vision*, pages 779–795, 2020.
- [36] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In *Advances in Neural Information Processing Systems*, 2021.

- [37] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *Proc. International Conference on Machine Learning*, pages 10596–10606, 2021.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019.
- [42] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, 2013.
- [43] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013.
- [44] Andrii Zadaianchuk, Matthias Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *International Conference on Learning Representations*, 2023.
- [45] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.
- [46] Runsheng Zhang, Yaping Huang, Mengyang Pu, Jian Zhang, Qingji Guan, Qi Zou, and Haibin Ling. Object discovery from a single unlabeled image by mining frequent itemsets with multi-scale features. *IEEE Transactions on Image Processing*, 29:8606–8621, 2020.
- [47] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814–2821, 2014.
- [48] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *Proc. European Conference on Computer Vision*, pages 391–405, 2014.
- [49] Hasib Zunair and A Ben Hamza. Learning to recognize occluded and small objects with partial inputs. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 675–684, 2024.