Unsupervised Camouflaged Object Segmentation as Domain Adaptation

Yi Zhang
LIVIA, École de Technologie Supérieure
Montreal, Canada

Chengyi Wu
Henan Polytechnic University
Henan, China

Abstract

*Deep learning for unsupervised image segmentation remains challenging due to the absence of human labels. The common idea is to train a segmentation head, with the supervision of pixel-wise pseudo-labels generated based on the representation of self-supervised backbones. By doing so, the model performance depends much on the distance between the distribution of target datasets, and the one of backbones' pre-training dataset (*e.g.*, ImageNet). In this work, we investigate a new task, namely unsupervised camouflaged object segmentation (UCOS), where the target objects own a common rarely-seen attribute,* i.e.*, camouflage. Unsurprisingly, we find that the state-of-the-art unsupervised models struggle in adapting UCOS, due to the domain gap between the properties of generic and camouflaged objects. To this end, we formulate the **UCOS** as a source-free unsupervised **d**omain **a**daptation task (**UCOS-DA**), where both source labels and target labels are absent during the whole model training process. Specifically, we define a source model consisting of self-supervised vision transformers pre-trained on ImageNet. On the other hand, the target domain includes a simple linear layer (*i.e.*, our target model) and unlabeled camouflaged objects. We then design a pipeline for foreground-background-contrastive self-adversarial domain adaptation, to achieve robust UCOS. As a result, our baseline model achieves superior segmentation performance when compared with competing unsupervised models on the UCOS benchmark, with the training set which's scale is only one tenth of the supervised COS counterpart. The UCOS benchmark and our baseline model are now publicly available*[1].

1. Introduction

In real-world scenes, there is a specific domain of objects which share one common attribute, namely "visual camouflage". Camouflaged objects introduce challenges to image segmentation with their different types of concealing coloration [9] (Figure 1). The common setting for camouflaged object segmentation (COS) is to fine-tune an encoder-decoder framework with well-labelled camouflaged objects [16, 55, 51, 71], based on the supervised ImageNet pre-trains [10, 21, 13]. Though improvement [71, 36] has been made as the booming development of vision transformers [13, 42], this setting requires either dense labels (*i.e.*, pixel-wise binary masks) or weak labels (*e.g.*, points, object categories) as the supervision for training COS models. To advance COS to open-world applications where extensive human labels are hardly gained, and supervised models tend to be poorly generalized [5, 31, 47]; We take advantage of self-supervised ImageNet-based pre-trains [5] and propose the first unsupervised COS baseline model, which requires no any human labels in the whole training pipeline.

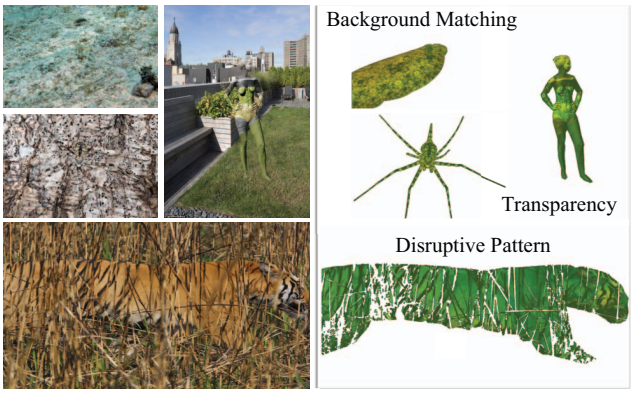

Figure 1: An illustration of camouflaged object segmentation. The camouflage domain-specific properties (*e.g.*, color-/texture-based background matching, transparency and disruptive pattern) are rarely-seen in generic object dataset such as ImageNet.

Intuitively, we formulate the **u**nsupervised **COS** as a task of source-free unsupervised **d**omain **a**daptation, abbreviated as **UCOS-DA** (Figure 2). Being different to com-

[1] https://github.com/Jun-Pu/UCOS-DA

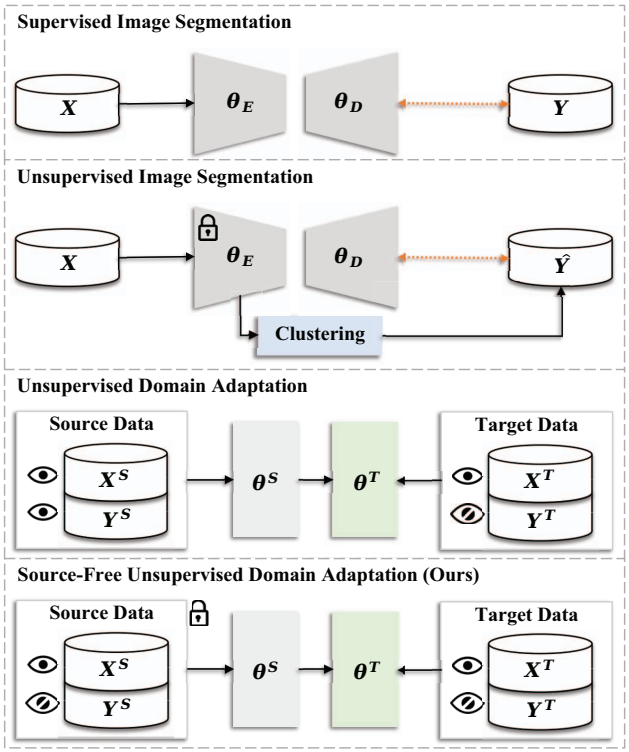

Figure 2: An illustration of related tasks. $\{X, Y, \hat{Y}\}$ denote images, ground truth and pseudo-labels generated by unsupervised backbones, respectively. $\{\theta^S, \theta^T\}$ indicate parameter sets of source and target models, respectively. $\{\theta_E, \theta_D\}$ means the parameter set of encoder and decoder of a given segmentation network. Note that in our task, namely UCOS-DA, the source model (θ^S) was trained in a self-supervised manner, without using source labels (Y^S).

mon source-free unsupervised domain adaptation settings where human labels are needed to train the source model, our UCOS-DA setting does not involve supervised training of the source domain. To conduct the new task, we propose a UCOS-DA baseline model consisting of three components, *i.e.*, a self-supervised source model, a light-weighted target model and an adversarial domain adaptation module (Figure 3). Following state-of-the-art unsupervised image segmentation methods [61, 38, 23, 48, 49, 43, 34], we use DINO [5]'s ImageNet pre-trained self-supervised vision transformer, as the unsupervised object-centric feature extractor (*i.e.*, our source model). Considering the ambiguity (Figure 1) between object parts and background region in COS, we explore to shift more attention to the local features representing the boundary of camouflaged targets, during domain adaptation. We thus design a self-adversarial training module to weight more importance to the boundary-specific object-centric representations. Meanwhile, the target model learns to segment camouflaged objects with the pixel-wise supervision of pseudo-labels gained from DINO features.

In a nutshell, by proposing the new task (UCOS-DA) in the context of fully unsupervised image segmentation, we investigate the domain transfer ability of state-of-the-art self-supervised vision transformers, especially towards the circumstance where a large discrepancy exists between the source domain and target domain (here we mean different visual patterns between generic objects and camouflaged objects). The main contributions are summarized as follows: **1)** We firstly investigate the task of unsupervised COS, by implementing a systematic benchmark study involving seven evaluative metrics and five state-of-the-art image segmentation methods. **2)** We investigate unsupervised COS from a perspective of source-free unsupervised domain adaptation, by proposing a baseline model which gains competitive results on multiple benchmark datasets. Besides, we discuss key issues for bridging domain adaptation to unsupervised object-centric representation learning. We hope our work could inspire more generalizable unsupervised image segmentation models in future researches.

2. Related Work

2.1. Self-Supervised Representation Learning

Learning to localize objects without using any human labels is a longstanding issue in the field of computer vision. The issue has recently appealed much more attention from the community, owing to the release of self-supervised representation learning methodologies, such as "MoCo Trilogy" [20, 7, 8], SimCLR [6], DenseCL [60], DINO [5], MAE [19] and "BEiT Trilogy" [4, 39, 57]. These models were trained with large-scale datasets (*e.g.*, ImageNet [10]) in a self-supervised manner, advancing the label-free object discovery. We briefly summarize recent self-supervised methods according to their types of pretext tasks:

Contrastive Learning. Pioneer works, MoCo [20] and SimCLR [6], proposed to optimize their networks' features via calculating similarities between two branches of features, respectively acquired from two sets of visual inputs. Notably, MoCo [20] used two encoders with different parameter updating strategies , while SimCLR [6] took advantage of one encoder with two sets of parameters (Siamese framework). Following MoCo, DenseCL [60] proposed dense projection heads to facilitate downstream unsupervised dense prediction tasks. Inspired by both MoCo and SimCLR, BYOL [17] used an on-line network and a target network to conduct self-supervised training, without relying on negative pairs. Following BYOL, DINO [5] applied two interactive encoders sharing the same ViT [13]-based architecture however with different parameter sets and updating strategies, achieved representations that illustrate superior object emergence when compared to the fully supervised

counterparts.

Masked Image Modeling (MIM). MIM-based methods aim to learn representations via reconstructing original images from image patches where a certain percentage of them are masked out. BeiT [4], as one of the pioneer works within this category, followed the masked language modeling strategy proposed in BERT [11] and introduced MIM into vision transformers. MAE [19] also proposed auto-encoder-like architecture but to reconstruct pixels rather than to predict tokens. BEiT-v2 [39] replaced the original reconstruction target with semantic-rich visual tokenizers to learn representations highlighting semantic cues. MaskFeat [62] also used MIM for model training however with the optimizing target of reconstructing HOG features of the masked image patches. SimMIM [65] proposed new prediction head consisting of only one linear payer.

Multi-Modal Alignment. The community recently witnessed a competition in establishing large vision-language models (VLMs) for representation learning [41, 66, 30, 52, 12, 57, 26]. Compared to vision-only self-supervised settings, VLMs relax the constrict of leveraging human labels by relying on image-text pairs, to learn multi-modal representations via aligning visual and textual cues. CLIP [41] jointly trained a text encoder and an image encoder to predict positive image-text pairs, achieving state-of-the-art zero-shot image classification. To further obtain object-centric locality-aware representation, GLIP [30] jointly optimized image and text encoders to localize positive region-word pairs. GroupViT [66] added grouping blocks to each level of a ViT [13], enabling progressive optimization of its vision encoder with only text-based weak supervision. Being different to above frameworks which rely on separate text and image encoders, CLIPPO [52] extracted both image and text features with a single encoder. Methods such as MaskCLIP [12] and BEiT-v3 [57] combined MIM strategy and visual-language contrastive learning to pursue generalizable representation. RO-ViT [26] achieved state-of-the-art open-vocabulary object detection, via manipulating ViT's positional embeddings at the pre-training stage and gaining region-aware image-text pairs at the fine-tuning stage. More recently, MUG [72] achieved new state-of-the-art in vision transfer learning tasks, via training a self-supervised vision-language model based on large-scale web data.

Despite the booming development of large-scale self-supervised multi-modal pre-trained models, unsupervised domain adaptation remains an open issue due to the finite scale of the pre-training data. To this end, OOD-CV [73] released an open challenge[2] to continually advance researches in exploring the transfer learning ability of state-of-the-art self-supervised pre-trained models.

[2] http://www.ood-cv.org/challenge.html

2.2. State-of-the-Art Unsupervised Segmentation

The "pre-training and fine-tuning" has been the most commonly-used paradigm for training deep neural networks since the emergence of ImageNet [10]. Recent development of self-supervised pre-trains (Section 2.1) stimulates the development of unsupervised image segmentation [18, 61, 38, 59, 70, 74, 69, 48, 23, 43, 49, 24, 58]. These methods are able to conduct instance-level pixel-wise classification without using any manual annotations.

Unsupervised Object Segmentation. TokenCut [61] conducted spectral clustering based on the DINO [5] features, yet the method is able to segment only one object per image. SelfMask [48] applied different number of clusters to produce multiple binary masks, and introduced a voting strategy to gain the final prediction. Also based on DINO features, FOUND [49] retrieved the background seed and identified its complement as the foreground. Final results were obtained by training a linear layer with the supervision of retrieved foreground masks. DINOSAUR [43] explored the task from a perspective of object-centric learning. The method was optimized to reconstruct the given images with slot-attention-based [33] decomposed object-centric representations. There is another class of methods [53, 1, 22, 3, 75] that use generative adversarial networks to generate the foreground masks representing target objects. Though progress was achieved during the past few years, we find that current unsupervised object segmentation methods tend to fail the cases where objects show complicated appearances in specific context (*e.g.*, camouflage, an object-centric attribute rarely-seen in ImageNet).

Unsupervised Semantic Segmentation. Thanks to the booming trend of large-scale self-supervised pre-trained models, the community witnessed an important change of the learning paradigm of semantic segmentation, from fully-/weakly-supervised learning to fully unsupervised learning. Recent methods such as STEGO [18], SpectralSeg [38], FreeSOLO [59], SelfPatch [70], TransFGU [69], Leopart [74], Odin [23], Exemplar-FreeSOLO [24] and CutLER [58], were trained to assign each pixel to specific object class without the supervision of any human labels. Similar to supervised methods, current unsupervised semantic segmentation methods face challenges such as occlusion detection, small object detection and multi-instance identification.

2.3. Unsupervised Domain Adaptation

Recent researches [27, 68, 64, 67, 40, 32, 54, 25, 44] investigated source-free unsupervised domain adaptation, where only the pre-trained source model and unlabeled target data are accessible during the domain adaptation. USFDA [27] proposed a source similarity metric to conduct domain adaptation without source data, and achieved on-par results when compared to the source-dependent coun-

terparts. G-SFDA [68] proposed local structure clustering to adapt source model to the target domain in the absence of source data. A^{2}Net [64] was trained to classify the target data into source-similar and source-dissimilar groups, via an adaptive adversarial strategy. NRC-SFDA [67] explored the local affinity of target data and achieved improved source-free adaptation upon both 2D and 3D target data. CPGA [40] disentangled the source model and gained class-wise features, namely avatar prototype, to facilitate source-target alignment. More recently, STPL [32] used temporal cues, *i.e.*, optical flow, to conduct domain adaptation for video semantic segmentation. ASFDA [54] resorted to active learning technique to identify a small set of source features, which supported the efficient training of the target model. C-SFDA [25] proposed new self-training strategy based on curriculum learning. MSFDA [44] explored multi-source-free domain adaptation and found an inherent bias-variance trade-off within the task, thus inspiring future works.

2.4. Uniqueness of Our Model

Training a segmentation head, with merely unlabeled COS dataset and ImageNet pre-trained self-supervised model, can be regarded as a source-free unsupervised domain adaptation task. Due to the out-of-distribution properties of camouflaged objects (Figure 1), unsupervised COS is an extremely challenging task. To this end, we consider to discovery and reserve the boundary-specific local self-supervised features, and resort to adversarial domain adaptation technique to improve the model transfer robustness. Besides the innovation towards the task formulation and camouflaged prior modeling, we define our target model as a simple linear layer yet predicts superior results when compared with its counterparts in unsupervised object segmentation.

3. UCOS-DA Methodology

We propose the first baseline model for **u**nsupervised **c**amouflaged **o**bject **s**egmentation, from the perspective of **d**omain **a**daptation (**UCOS-DA**). The model consists of a self-supervised ImageNet-based pre-trained vision transformer as the source model (θ^S), a linear-probe layer as the target model (θ^T), and a **f**oreground-**b**ackground-contrastive self-adversarial domain **a**daptation module (θ^D, abbreviated as **FBA**). The pipeline of the proposed baseline model is shown in Figure 3.

3.1. UCOS-DA Motivation&Formulation

A popular chatbot gives a definition towards "object": "*An object refers to a distinct item or entity that occupies space, has properties, can be perceived through our senses*". In 2D domain, object segmentation (*a.k.a.*, object-level pixel-wise classification) models usually require manual annotations as supervision to learn the mapping from images to objects. As the recent development of self-supervised models, it is inspiring to see that, specific pretext tasks (*e.g.*, enforcing the view-invariance [17, 5], recovering the missing parts [19]), enable deep learning models to discover object concepts without using external supervision of human labels. Thus, self-supervised learning seems to be a more humanoid learning paradigm and thus promising. In the context of unsupervised COS, we aim to achieve a model which learns the camouflaged properties with only unlabelled image data, thus segmenting objects concealed in various real-world scenes effectively. Considering the absence of large-scale camouflage pre-trains, a feasible solution is to extract features from generic data-based self-supervised models and adapt them to the camouflage domain.

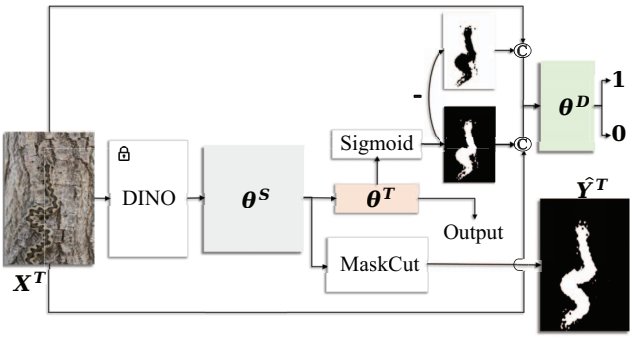

Figure 3: The pipeline of our proposed UCOS-DA baseline model. The model consists of a frozen source model (θ^S), a light-weighted linear target model (θ^T) and a foreground-background-contrastive self-adversarial domain adaptation module (θ^D). Notably, no any human labels are used for UCOS-DA pseudo-labelling, pre-training or fine-tuning.

To this end, we formulate the objective of UCOS-DA as minimizing an empirical loss function:

$$\begin{aligned}\min_{\{\theta^S,\theta^D,\theta^T\}} & \mathbb{E}_{X^T,\hat{Y}^T}[\mathcal{L}(f^T(X^T;\theta^S,\theta^D,\theta^T),\hat{Y}^T)] \\ &= \int \mathcal{L}(f^T(X^T;\theta^S,\theta^D,\theta^T),\hat{Y}^T)dp(X^T,\hat{Y}^T) \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(f^T(x_i^T;\theta^S,\theta^D,\theta^T),\hat{y_i}^T),\end{aligned} \tag{1}$$

with

$$(x_i^T,\hat{y_i}^T) \sim p(X^T,\hat{Y}^T), \tag{2}$$

where $\{\theta^S,\theta^D,\theta^T\}$ denotes parameter sets of the source model, the FBA module and the target model, respectively. $(x_i^T,\hat{y}_i^T)$ denotes a sample pair from the joint data distribution in the target domain. Notably, the $\hat{Y}^T$ indicates the pseudo-labels corresponding to the training data in the target domain. $\mathcal{L}(\cdot)$ means the loss function.

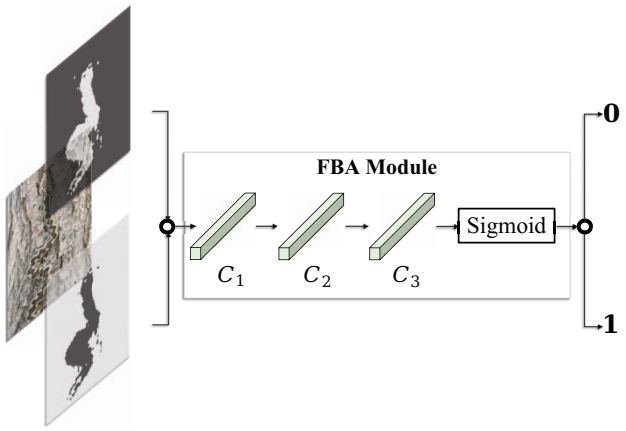

Figure 4: The architecture of the FBA (**f**oreground-**b**ackground-contrastive self-adversarial domain **a**daptation) module (θ^D). $\{C_1, C_2, C_3\}$ denotes the number of channels of each linear layer, respectively.

3.2. UCOS-DA Architecture

Generic Object-Centric Knowledge Extraction. Acorrding to previous researches [61, 49, 58], among various self-supervised pre-trains, DINO [5] has proved its superior object emergence ability and is regarded as one of most promising candidates for downstream unsupervised image segmentation tasks. We use DINO ImageNet pretrains as our source model, and extract its rich generic object knowledge to generate pseudo-labels, and to facilitate a self-supervised training of the target model.

Pseudo-Labels. We resort to normalized cuts technique [45] to generate coarse maps based on DINO features. Specifically, we resort to MaskCut methodology [58], which conducts multiple iterations of normalized cuts with DINO features, based on a patch-level affinity matrix.

Adversarial Domain Adaptation. To adapt DINO pretrained features to unsupervised COS, we first study the object priors when it comes to the camouflaged scenario. In fact, animals tend to deceive predators' visual perception with specific concealing coloration. As a consequence, noisy visual cues are brought to the boundary region of camouflaged objects in 2D images, making it hard to obtain on-par segmentation results. We argue that the blur of object boundary is one of the main cause for the big divergence between camouflaged and generic data distributions.

To close the gap between the source (generic) domain and the target (camouflage) domain, we suggest a new module to emphasize the reservation of boundary-specific local representations of source model, during training the target model. Specifically, we introduce a foreground-background-contrastive self-adversarial domain adaptation (FBA) module (Figure 4), to conduct a sub-task aiming at further distinguishing the predicted foreground maps from their complements. Our FBA module mainly consists of three hierarchical linear layers, computing foreground/background class score ($S, S \in [0, 1]$) as:

$$S = \sigma(FC_{C3}(LR(FC_{C2}(LR(FC_{C1}(Cat(X, P'))))))), \tag{3}$$

where $\{X, P'\}$ means the given images and corresponding binary masks gained via the target model. $\sigma(\cdot)$, $FC(\cdot)$, $LR(\cdot)$ and $Cat(\cdot)$ denotes Sigmoid function, linear(fully-connected) layer, leakyReLu activation layer and concatenation operation, respectively.

3.3. Implementation Details

Loss Function. As the target model and FBA module are co-trained for domain adaptation, the total loss ($\mathcal{L}$) of our UCOS-DA baseline model is thus formulated as the sum of a segmenting loss ($\mathcal{L}^{Seg.}$) and an adversarial loss ($\mathcal{L}^{Adv.}$):

$$\mathcal{L} = \mathcal{L}^{Seg.}(P, \hat{Y}^T) + \mathcal{L}^{Adv.}(S, C), \tag{4}$$

where P and C ($C \in \{0, 1\}$) denotes segmentation results of target model, and foreground/background class label, respectively. Notably, in this work, we apply the structure loss [63] as the segmentation loss $\mathcal{L}^{Seg.}$, and binary cross entropy loss as the foreground/background classification loss (*i.e.*, adversarial loss $\mathcal{L}^{Adv.}$).

Hyper-Parameters. We train the UCOS-DA baseline model by using PyTorch with a maximum epoch of 5. The images are re-scaled to the size of 512×512 during training. The initial learning rate of the target model and the FBA module is set to 5e-3 and 5e-4, respectively.

4. Experiments

4.1. Settings

Training DataSets. We randomly collect 300 images from the most commonly-used supervised COS training set [16, 71], which includes 4,040 images representing various camouflage-based scenes. We also randomly select 300 images from the most commonly-used salient object segmentation training set, *i.e.*, DUTS-tr [56]. Thus, the training set for our UCOS-DA consists of 600 images covering wide real-world daily scenes, while has its scale much smaller than the ones for fully-supervised image segmentation.

Testing DataSets. To thoroughly analyze the performance of our new unsupervised baseline, we test our model and all benchmark models on six commonly-used testing sets, *i.e.*, ECSSD [46], HKU-IS [29], CAMO [28], CHAMELEON [50], COD10K [16] and NC4K [35], which possess 1K, 4447, 250, 76, 2026 and 4121 images, respectively.

Benchmark Models. To contribute the community a comprehensive benchmark towards unsupervised object segmentation, we collect most recent state-of-the-art fully unsupervised models, including BigGW [53], TokenCut [61], SpectralSeg [38], SelfMask [48] and FOUND [49].

Table 1: Comparison of our UCOS-DA and state-of-the-art unsupervised methods on salient object segmentation benchmark datasets. The **best** and the second best results of each row are highlighted.

Task	Dataset	Metric	BigGW ICML'21 [53]	TokenCut CVPR'22 [61]	TokenCut w/ B.S. CVPR'22 [61]	SpectralSeg CVPR'22 [38]	SelfMask CVPRw'22 [48]	SelfMask w/ U.B. CVPRw'22 [48]	FOUND CVPR'23 [49]	UCOS-DA(**Ours**) ICCVw'23
Salient Object Segmentation	ECSSD [46]	mIoU ↑	.689	.712	.774	.733	.779	.787	.805	**.816**
		Acc. ↑	.905	.918	.934	.891	.943	.946	.948	**.951**
		F_β^{max} ↑	.800	.803	.874	.805	.892	**.897**	.896	.891
		F_β^{mean} ↑	.654	.801	.714	.803	.861	.867	**.894**	.888
		F_β^W ↑	.568	.785	.630	.790	.846	.852	**.877**	.876
		S_α ↑	.783	.807	.832	.806	.866	.871	.875	**.878**
		E_ϕ^{max} ↑	.871	.886	.905	.865	.928	.932	.932	**.934**
		E_ϕ^{mean} ↑	.714	.884	.755	.862	.920	.925	.930	**.931**
		$\mathcal{M}$ ↓	.169	.082	.129	.109	.058	.055	.052	**.049**
	HKU-IS [29]	mIoU ↑	.641	.608	.673	.735	.747	.755	.787	**.794**
		Acc. ↑	.905	.916	.936	.932	.949	.951	.958	**.959**
		F_β^{max} ↑	.760	.741	.832	.815	.869	.874	**.877**	.872
		F_β^{mean} ↑	.611	.739	.667	.812	.830	.836	**.875**	.870
		F_β^W ↑	.515	.703	.557	.801	.818	.824	**.863**	.861
		S_α ↑	.761	.748	.777	.828	.851	.856	.869	**.871**
		E_ϕ^{max} ↑	.859	.866	.871	.896	.930	.934	**.939**	.937
		E_ϕ^{mean} ↑	.696	.864	.728	.894	.919	.923	**.936**	.935
		$\mathcal{M}$ ↓	.166	.084	.123	.068	.052	.050	.042	**.041**

Table 2: Comparison of our UCOS-DA baseline and state-of-the-art unsupervised methods on camouflaged object segmentation benchmark datasets. The **best** and the second best results of each row are highlighted.

Task	Dataset	Metric	BigGW ICML'21 [53]	TokenCut CVPR'22 [61]	TokenCut w/ B.S. CVPR'22 [61]	SpectralSeg CVPR'22 [38]	SelfMask CVPRw'22 [48]	SelfMask w/ U.B. CVPRw'22 [48]	FOUND CVPR'23 [49]	UCOS-DA(**Ours**) ICCVw'23
Camouflaged Object Segmentation	CAMO [28]	mIoU ↑	.322	.431	.422	.411	.418	.430	.505	**.528**
		Acc. ↑	.775	.837	.838	.765	.813	.819	.871	**.873**
		F_β^{max} ↑	.428	.546	.550	.486	.549	.561	.635	**.647**
		F_β^{mean} ↑	.349	.543	.434	.481	.536	.547	.633	**.646**
		F_β^W ↑	.299	.498	.383	.450	.483	.495	.584	**.606**
		S_α ↑	.565	.633	.639	.579	.617	.627	.685	**.701**
		E_ϕ^{max} ↑	.678	.708	.699	.658	.713	.724	.784	**.786**
		E_ϕ^{mean} ↑	.528	.706	.595	.648	.698	.708	.782	**.784**
		$\mathcal{M}$ ↓	.282	.163	.195	.235	.188	.182	.129	**.127**
	CHAMELEON [50]	mIoU ↑	.267	.436	.415	.381	.396	.406	.468	**.525**
		Acc. ↑	.807	.868	.871	.780	.825	.832	**.905**	**.905**
		F_β^{max} ↑	.356	.540	.544	.446	.511	.522	.591	**.631**
		F_β^{mean} ↑	.294	.536	.393	.440	.481	.491	.590	**.629**
		F_β^W ↑	.244	.496	.351	.410	.436	.447	.542	**.591**
		S_α ↑	.547	.654	.655	.575	.619	.629	.684	**.715**
		E_ϕ^{max} ↑	.662	.743	.734	.638	.726	.734	**.812**	.804
		E_ϕ^{mean} ↑	.527	.740	.582	.628	.675	.683	**.810**	.802
		$\mathcal{M}$ ↓	.257	.132	.169	.220	.176	.169	**.095**	**.095**
	COD10K [16]	mIoU ↑	.236	.415	.423	.331	.388	.397	.428	**.462**
		Acc. ↑	.798	.897	.903	.807	.870	.875	**.915**	.914
		F_β^{max} ↑	.315	.509	.537	.395	.504	.514	.521	**.548**
		F_β^{mean} ↑	.246	.502	.399	.388	.469	.478	.520	**.546**
		F_β^W ↑	.185	.469	.334	.360	.431	.440	.482	**.513**
		S_α ↑	.528	.658	.666	.575	.637	.645	.670	**.689**
		E_ϕ^{max} ↑	.670	.740	.739	.606	.718	.728	**.753**	.741
		E_ϕ^{mean} ↑	.497	.735	.609	.595	.679	.687	**.751**	.740
		$\mathcal{M}$ ↓	.261	.103	.127	.193	.131	.125	**.085**	.086
	NC4K [35]	mIoU ↑	.382	.546	.561	.495	.529	.538	.566	**.590**
		Acc. ↑	.814	.899	.904	.841	.887	.891	**.916**	.915
		F_β^{max} ↑	.484	.655	.682	.570	.661	.670	.676	**.691**
		F_β^{mean} ↑	.391	.649	.547	.562	.634	.642	.674	**.689**
		F_β^W ↑	.319	.615	.478	.535	.593	.601	.637	**.656**
		S_α ↑	.608	.725	.735	.669	.716	.723	.741	**.755**
		E_ϕ^{max} ↑	.714	.806	.807	.729	.796	.803	**.827**	.822
		E_ϕ^{mean} ↑	.565	.802	.683	.719	.777	.784	**.824**	.819
		$\mathcal{M}$ ↓	.246	.101	.133	.159	.114	.110	**.084**	.085

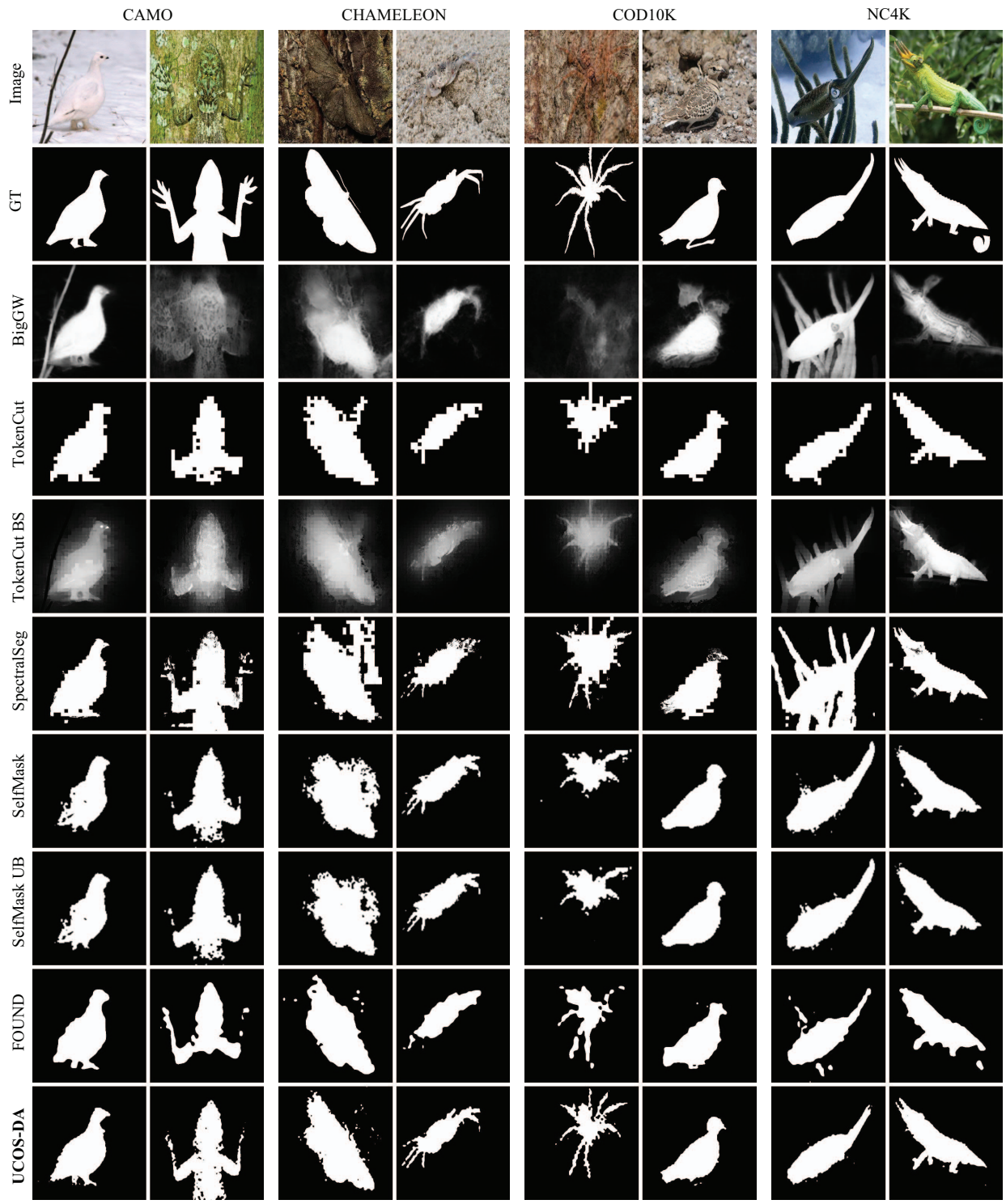

Figure 5: Visual samples of our baseline model (UCOS-DA) and all competing models.

Evaluation Metrics. We apply seven widely-used metrics to quantitatively evaluate all the benchmark models. The metrics include Accuracy ($Acc.$), mean Intersection over Union ($mIoU$), mean absolute error (M), F-measure [2]

Table 3: Comparison of linear-probe strategies upon COS.

Method	COD10K			NC4K		
	mIoU ↑	Acc. ↑	F_β^{max} ↑	mIoU ↑	Acc. ↑	F_β^{max} ↑
FOUND	42.8	**91.5**	52.1	56.6	**91.6**	67.6
FOUND (F.T.)	-4.0	-1.8	-4.8	-3.8	-1.5	-4.6
Ours	**46.2**	91.4	**54.8**	**59.0**	91.5	**69.1**

(F_β), weighted F-measure [37] (F_β^W), S-measure [14] (S_α) and E-measure [15] (E_ϕ). Notably,
F_β computes both $Precision$ and $Recall$, formulated as:

$$F_\beta = \frac{(1+\beta^2) Precision\ Recall}{\beta^2 Precision + Recall}, \tag{5}$$

with

$$Precision = \frac{|P \cap G|}{|P|}; Recall = \frac{|P \cap G|}{|G|}, \tag{6}$$

where G is the ground truth and P denotes a binarized predictions. Multiple P are computed by assigning different integral thresholds τ ($\tau \in [0, 255]$) to the predicted map. The β^2 is commonly set to 0.3 .
S_α evaluates the structural similarities between the prediction and the ground truth. The metric is defined as:

$$S = \alpha S_o + (1-\alpha) S_r, \tag{7}$$

where S_r and S_o denote the region-/object-based structure similarities, respectively. $\alpha \in [0, 1]$ is empirically set as 0.5 to arrange equal weights to both region-level and object-level quantitative evaluation.

4.2. Comparison with Unsupervised Methods

Zero-Shot Transfer. As shown in Table 1 and Table 2, we benchmark all competing models on datasets for both camouflaged and salient object segmentation. As a result, our UCOS-DA baseline model obtains overall superior performance on multiple testing sets. Please note that the benchmark results are all based on the codes and released checkpoints from each model's official project page. We also show some visual samples in Figure 5.
Linear Probe via Adversarial training. To analyze the effectiveness of our proposed FBA module, we compare our results with the ones of FOUND [49], which also uses linear probe-based DINO fine-tuning strategy. As a result, we spot a slight performance drop of FOUND when fine-tuning its linear layer with COS training set (Table 3). We also show a visual example to further illustrate the phenomenon (Figure 6). On the contrary, our method not only performs superior results on COS testing sets, but also acquires competitive results on salient object segmentation datasets, indicating the effectiveness and robustness of the proposed modules.

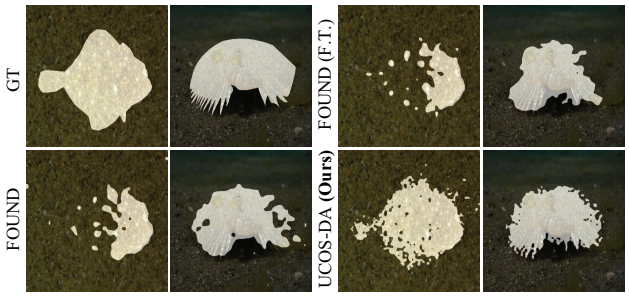

Figure 6: Visual comparison of different linear-probe-based unsupervised image segmentation methods.

5. Conclusion and Future Work

In this work, we investigate a new challenging image segmentation task, *i.e.*, unsupervised camouflaged object segmentation. We firstly contribute a comprehensive benchmark study to show limited transferring ability of state-of-the-art unsupervised image segmentation models. We explore the co-existence of challenge and opportunity of a unique object-centric attribute, *i.e.*, concealing coloration, and resort to prior-inspired adversarial domain adaptation to conduct the task. As a result, our new baseline model achieves overall superior scores based on multiple metrics and testing sets. Based on our study towards UCOS-DA, we find following issues that could be paid attention in the future researches.
Attribute-based Domain Adaptation. The concealing coloration attribute makes unsupervised COS an open issue in both societies of unsupervised domain adaptation and unsupervised image segmentation, in the context of current generic object datasets-based pre-trains. Future works may explore more towards specific domains where the objects own rarely-seen attributes, and investigate attribute-specific domain adaptation methods.
Generalizability of Self-Supervised Pre-trains. Our benchmark shows limited application of current self-supervised pre-trained models, which could inspire more studies towards generalizable pre-trains. Investigating the transfer learning ability of self-supervised pre-trained models is essential, since it is expensive to train large models in each domain. Besides, exploring effective domain adaptation methods under challenging settings helps to advance the development of interpretable AI. We hope our preliminary work could inspire future researches towards more generalizable label-free segmentation and unsupervised domain adaptation methodologies.
Other Learning Paradigms. Besides "pre-training and fine-tuning", future researches may explore unsupervised representation decomposition with attribute-sufficient real-world data, aiming to acquire both interpretability and generalizability.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In *IEEE International Conference on Computer Vision (ICCV)*, pages 13970–13979, 2021. 3

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1597–1604, 2009. 7

[3] Jeongmin Bae, Mingi Kwon, and Youngjung Uh. Furrygan: High quality foreground-aware image synthesis. *European Conference on Computer Vision (ECCV)*, 2022. 3

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1, 2, 3, 4, 5

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pages 1597–1607. PMLR, 2020. 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 2

[9] Hugh Bamford Cott. Adaptive coloration in animals. 1940. 1

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 1, 2, 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[12] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10995–11005, 2023. 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3

[14] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4548–4557, 2017. 8

[15] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 2021. 8

[16] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2777–2787, 2020. 1, 5, 6

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, 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. *Advances in neural information processing systems (NeurIPS)*, 33:21271–21284, 2020. 2, 4

[18] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *International Conference on Learning Representations (ICLR)*, 2022. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 2, 3, 4

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 9729–9738, 2020. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[22] Xingzhe He, Bastian Wandt, and Helge Rhodin. Ganseg: Learning to segment by unsupervised hierarchical image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1225–1235, 2022. 3

[23] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *European Conference on Computer Vision (ECCV)*, pages 123–143. Springer, 2022. 2, 3

[24] Taoseef Ishtiak, Qing En, and Yuhong Guo. Exemplar-freesolo: Enhancing unsupervised instance segmentation with exemplars. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15424–15433, June 2023. 3

[25] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24120–24131, 2023. 3, 4

[26] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11144–11154, 2023. 3

[27] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 4544–4553, 2020. 3

[28] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding (CVIU)*, 184:45–56, 2019. 5, 6

[29] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5455–5463, 2015. 5, 6

[30] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, 2022. 3

[31] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. *International Conference on Learning Representations (ICLR)*, 2022. 1

[32] Shao-Yuan Lo, Poojan Oza, Sumanth Chennupati, Alejandro Galindo, and Vishal M Patel. Spatio-temporal pixel-level contrastive learning-based source-free domain adaptation for video semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10534–10543, 2023. 3, 4

[33] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:11525–11538, 2020. 3

[34] Yunqiu Lv, Jing Zhang, Nick Barnes, and Yuchao Dai. Weakly-supervised contrastive learning for unsupervised object discovery. *arXiv preprint arXiv:2307.03376*, 2023. 2

[35] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 11591–11601, 2021. 5, 6

[36] Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv, Xinyu Tian, Deng-Ping Fan, and Nick Barnes. Transformer transforms salient object detection and camouflaged object detection. *arXiv preprint arXiv:2104.10127*, 1(2):5, 2021. 1

[37] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255, 2014. 8

[38] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8364–8375, 2022. 2, 3, 5, 6

[39] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2, 3

[40] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021. 3, 4

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021. 3

[42] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 1

[43] 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. *International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[44] Maohao Shen, Yuheng Bu, and Gregory W Wornell. On balancing bias and variance in unsupervised multi-source-free domain adaptation. In *International Conference on Machine Learning (ICML)*, pages 30976–30991. PMLR, 2023. 3, 4

[45] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence (TPAMI)*, 22(8):888–905, 2000. 5

[46] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 38(4):717–729, 2015. 5, 6

[47] Yuge Shi, Imant Daunhawer, Julia E Vogt, Philip Torr, and Amartya Sanyal. How robust is unsupervised representation learning to distribution shift? In *International Conference on Learning Representations (ICLR)*, 2023. 1

[48] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3971–3980, June 2022. 2, 3, 5, 6

[49] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonin Vobecky, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5, 6, 8

[50] Przemysław Skurowski, Hassan Abdulameer, Jakub Baszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozie. Animal camouflage analysis: Chameleon database. In *Unpublished Manuscript*, 2018. 5, 6

[51] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1025–1031, 2021. 1

[52] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels

only. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11006–11017, 2023. 3

[53] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *International Conference on Machine Learning (ICML)*, pages 10596–10606. PMLR, 2021. 3, 5, 6

[54] Fan Wang, Zhongyi Han, Zhiyan Zhang, Rundong He, and Yilong Yin. Mhpl: Minimum happy points learning for active source free domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20008–20018, 2023. 3, 4

[55] Kang Wang, Hongbo Bi, Yi Zhang, Cong Zhang, Ziqi Liu, and Shuang Zheng. D^2c-net: A dual-branch, dual-guidance and cross-refine network for camouflaged object detection. *IEEE Transactions on Industrial Electronics (TIE)*, 2021. 1

[56] 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 *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 136–145, 2017. 5

[57] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186, 2023. 2, 3

[58] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3124–3134, 2023. 3, 5

[59] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14176–14186, 2022. 3

[60] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3024–3033, 2021. 2

[61] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14543–14553, 2022. 2, 3, 5, 6

[62] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14668–14678, 2022. 3

[63] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI conference on artificial intelligence (AAAI)*, pages 12321–12328, 2020. 5

[64] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9010–9019, 2021. 3, 4

[65] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 3

[66] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18134–18144, 2022. 3

[67] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems (NeurIPS)*, 34:29393–29405, 2021. 3, 4

[68] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8978–8987, 2021. 3, 4

[69] Zhaoyuan Yin, Pichao Wang, Fan Wang, Xianzhe Xu, Hanling Zhang, Hao Li, and Rong Jin. Transfgu: A top-down approach to fine-grained unsupervised semantic segmentation. *European Conference on Computer Vision (ECCV)*, 2022. 3

[70] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8354–8363, 2022. 3

[71] Yi Zhang, Jing Zhang, Wassim Hamidouche, and Olivier Deforges. Predictive uncertainty estimation for camouflaged object detection. *IEEE Transactions on Image Processing (TIP)*, 2023. 1, 5

[72] Bingchen Zhao, Quan Cui, Hao Wu, Osamu Yoshie, and Cheng Yang. Vision learners meet web image-text pairs. *arXiv preprint arXiv:2301.07088*, 2023. 3

[73] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: a benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European Conference on Computer Vision (ECCV)*, pages 163–180. Springer, 2022. 3

[74] Adrian Ziegler and Yuki M Asano. Self-supervised learning of object parts for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14502–14511, 2022. 3

[75] Qiran Zou, Yu Yang, Wing Yin Cheung, Chang Liu, and Xiangyang Ji. Ilsgan: Independent layer synthesis for unsupervised foreground-background segmentation. *The Association for the Advancement of Artificial Intelligence (AAAI)*, 2023. 3