LEARNING VIDEO CAMOUFLAGED OBJECT DETECTION FROM UNLABELED STATIC IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Camouflaged Object Detection (VCOD) is a criuital security technology for identifying camouflaged objects in videos. Existing VCOD methods often rely on motion signals across consecutive frames to capture dynamic cues. However, these motion signals are inherently noisy and inaccurate, and excessive reliance on them can lead to error accumulation. Moreover, the lack of camouflaged video datasets and the costly annotation process hinder progress in this field. To address these challenges, we propose SDC, a static-dynamic correspondence framework that extracts pseudo-dynamic signals from unlabeled static images, effectively capturing static-dynamic correspondences. Initially, we leverage coordinate data from the contrastive crop views to establish static correspondences, ensuring the consistency of static feature representations. Following that, we design a dynamic capture layer that generates pseudo-dynamic signals in both forward and backward directions. These pseudo-dynamic signals serve as cues for learning dynamic representations. Finally, we introduce a static-dynamic consistency loss function to enforce consistency between static and dynamic correspondences. Notably, our method requires only unlabeled static image data for training, freeing VCOD from its dependency on video data. By using only static images for training, SDC outperforms existing video-dependent strategies, surpassing state-of-the-art weakly supervised, unsupervised, and even surpasses several fully supervised video methods on public benchmarks.

1 INTRODUCTION

Video Camouflaged Object Detection (VCOD) is an important task in intelligent video content understanding, aiming to detect camouflaged objects that blend with their environment. It plays a critical role in real-world vision applications such as surveillance (Zhong et al., 2024), autonomous driving (Fan et al., 2022a), and ecological monitoring (Chen et al., 2025). Existing VCOD methods primarily rely on explicit motion cues extracted through optical flow (Zhang et al., 2025b) (as shown in Figure 1) or utilize temporal modeling architectures (Hui et al., 2024b) to capture implicit motion. However, research by Cheng et al. indicates that the high similarity between camouflaged objects and their backgrounds impairs the accurate estimation of optical flow and motion information in videos (Cheng et al., 2022a). As a result, the accumulation of incorrect motion signals can severely mislead the model's learning process, diminishing detection performance. Moreover, the scarcity of camouflaged video datasets significantly hinders the progress of VCOD technology. Existing methods rely entirely on video datasets for training (Lu et al., 2025), and these methods typically use weakly supervised, fully supervised, or semi-supervised annotations. This reliance on annotated video data makes these approaches not only data-intensive but also expensive and time-consuming, limiting their widespread applicability and slowing down the advancement of the field.

We observe that in natural vision, when humans view a static scene, the brain processes spatial and appearance information to form an initial understanding of the environment (Steel et al., 2021). This enables us to recognize objects and their surroundings, even in the absence of any movement or dynamic changes. However, when certain parts of the scene are altered or rearranged, the brain uses prior spatial knowledge along with new visual cues to perceive relative motion between these elements, as if the static scene were undergoing dynamic changes. Given this, a crucial question arises: if the human brain can detect and adapt to changes by leveraging spatial and motion-sensitive mechanisms, even in the absence of actual motion, can we also apply similar principles in machine

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

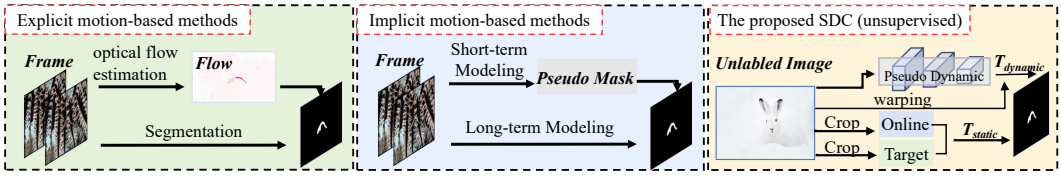
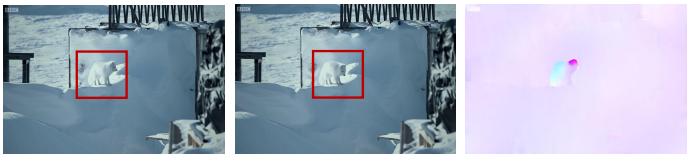


Figure 2: Different strategies of motion handling in VCOD: (a) Explicit motion-based methods, where optical flow maps are directly fed to the model for segmentation; (b) Implicit motion-based methods that learn motion cues through temporal modeling architectures, including short-term and long-term modeling techniques; (c) The proposed SDC method, which uses pseudo-dynamic signals from static images to learn static and dynamic consistency, without video data or annotations.

learning? Specifically, can we use only unlabeled static images to learn spatiotemporal correspondences (Jabri et al., 2020) and achieve more effective VCOD? By leveraging the spatial consistency in static images and simulating dynamic motion cues, we aim to bridge the gap between static and dynamic features. This would allow us to build a VCOD system without requiring pixel-level annotated video datasets, thus addressing the challenges posed by the scarcity of labeled video data.



(a) Frame T (b) Frame $T+1$ (C) Optical Flow

Figure 1: Two adjacent frames showing a camouflaged object highlighted by a bounding box.

To address this issue, we propose a method called Static-Dynamic Correspondence Network (SDC), which integrates static and dynamic correspondence learning into an efficient VCOD framework. Unlike traditional video data-driven methods, SDC eliminates the complexity of cross-frame reconstruction while achieving superior performance and resource efficiency, without requiring any video data. By specifically leveraging unlabeled static images (COD datasets), our method utilizes a large number of unlabeled images that depict diverse scenes, objects, and perspectives, forming a powerful and widely applicable visual data repository for the development of VCOD models. However, a key challenge is that static images lack dynamic signals, which are crucial features for the VCOD task. Injecting dynamic signals into static images is a non-trivial task.

To tackle this challenge, we propose a pseudo-dynamic learning strategy aimed at extracting pseudo-dynamic information from unlabeled static images while preserving the inherent static feature within these images. Our model learns two key aspects: (i) static consistency of cropped image views; (ii) dynamic consistency of pseudo-dynamic signals. By modeling the similarity between cropped views, we achieve static consistency, tightly aligning pixel pairs in overlapping regions. The introduction of pseudo-dynamic signals ensures dynamic consistency, which is achieved through a specially designed pseudo-dynamic generation network. This network is responsible for calculating pixel offset vectors between overlapping regions of cropped image views. These offset vectors represent the displacement of corresponding pixels in different cropped views, simulating the dynamic motion that would typically occur in video.

To construct positive masks, we leverage the coordinate relationships between cropped views. These masks play a crucial role in identifying regions with deterministic positive pixel pairs within both the static image (static) and pseudo-dynamic signals (dynamic). In this framework, our method ensures that the representations learned solely from images achieve static-dynamic consistency, enabling us to reach the desired goal. The contributions of this work are as follows: **1)** We propose the annotation-free VCOD framework, which enables cross-modal learning by integrating static and dynamic correspondence and learning from static images for video processing. **2)** We propose a novel motion handling strategy for VCOD, where pseudo-dynamic signals are extracted from static images to simulate dynamic motion, allowing effective motion handling without relying on video data, as shown in Figure 2. **3)** Our method requires only unlabeled static image data during training, freeing VCOD from video data dependency, and outperforms existing video-dependent supervised strategies. Notably, our model achieves high performance with a total training time of just one hour on a RTX4090 GPU. **4)** Our model achieves state-of-the-art performance on several public VCOD datasets, demonstrating the effectiveness of the proposed method.

2 RELATED WORKS

COD. Camouflaged object detection aims to segment objects whose appearances closely resemble their surroundings. Early deep COD methods mainly improve foreground–background discrimination through multi-scale representation, progressive refinement, and auxiliary supervision. For example, ZoomNet (Pang et al., 2022) exploits mixed-scale representations to capture camouflaged objects with large appearance and scale variations, while SINet-V2 (Fan et al., 2022b) progressively searches for and identifies concealed regions through refined feature interactions. More recently, research has increasingly explored annotation-efficient and foundation-model-based solutions. EASE (Du et al., 2025) introduces an environment-aware unsupervised framework based on environment prototype retrieval. Chen et al. (Chen et al., 2026) further leverage multimodal large language models and graded knowledge distillation for unsupervised COD. These advances substantially improve static-image COD, but do not explicitly address temporal correspondence and motion information required in video scenarios.

VCOD. For VCOD, modeling motion and temporal cues is essential. Existing methods mainly follow two directions (Zhang et al., 2025a). The first relies on optical flow. Ji et al. (Ji et al., 2021) refined flow-based motion features to reduce noise during multi-frame propagation, while Yan et al. (Yan et al., 2019) introduced recurrent cross-refinement to alleviate long-term error accumulation, at the cost of higher computational complexity. The second direction adopts implicit temporal modeling. Cheng et al. (Cheng et al., 2022b) used spatiotemporal transformers, and Yang et al. (Yang et al., 2021) combined short-term motion with long-range context through multi-granularity fusion. More recently, Xiao et al. (Xiao et al., 2026) explored self-supervised VCOD by disentangling motion and contextual information, reducing the dependence on dense annotations. However, existing approaches still learn temporal representations directly from video sequences. In contrast, we propose an annotation-free VCOD framework that synthesizes static and dynamic correspondences from unlabeled static images to emulate motion cues without requiring videos during training.

3 METHOD

In this study, we present an annotation-free method for VCOD using unlabeled static images. Our method constructs pseudo-dynamic cues from overlapping cropped views of static images to learn static and pseudo-dynamic correspondences. We show the architecture of the SDC in Fig. 3.

Framework. Given an input image $\boldsymbol{I}$, our framework first generates two distinct yet semantically related views, denoted as $\boldsymbol{I}_1$ and $\boldsymbol{I}_2 \in \mathbb{R}^{3\times h\times w}$. This is achieved using the ContrastiveCrop transformations (Peng et al., 2022), which have been specifically designed for contrastive learning. The core motivation behind ContrastiveCrop is to ensure that generated crops capture diverse local regions while remaining semantically consistent. Unlike naive random cropping, which can easily result in trivial or uninformative patches dominated by background pixels, ContrastiveCrop integrates semantic-aware localization with center-suppressed sampling to produce high-quality positive pairs. This ensures that both crops are object-centric, informative, and diverse enough to provide complementary learning signals for robust representation learning.

Semantic-aware localization. Semantic-aware localization identifies the object of interest within $\boldsymbol{I}$ using the heatmap M derived from intermediate feature activations. A bounding box B enclosing the salient regions is computed as:

$$B = L(\mathbb{I}[M > k]), \quad k \in [0, 1], \tag{1}$$

where k denotes a confidence threshold, $\mathbb{I}[\cdot]$ is the indicator function, and $L(\cdot)$ extracts the bounding region. By applying thresholding on the heatmap, this step adaptively localizes semantically meaningful regions instead of relying on uniform random sampling. The crop center (x, y) is then sampled from the bounding box B using a center-suppressed sampling strategy. Unlike naive central sampling, this method avoids redundancy and increases variability. Specifically, the coordinates are drawn from a Beta distribution $\mathrm{Beta}(\alpha, \alpha)$, where the hyperparameter $\alpha \in (0, 1)$ controls the degree of center suppression:

$$\begin{aligned} x &= B_{x0} + (B_{x1} - B_{x0}) \cdot u, \quad u \sim \mathrm{Beta}(\alpha, \alpha), \\ y &= B_{y0} + (B_{y1} - B_{y0}) \cdot v, \quad v \sim \mathrm{Beta}(\alpha, \alpha). \end{aligned} \tag{2}$$

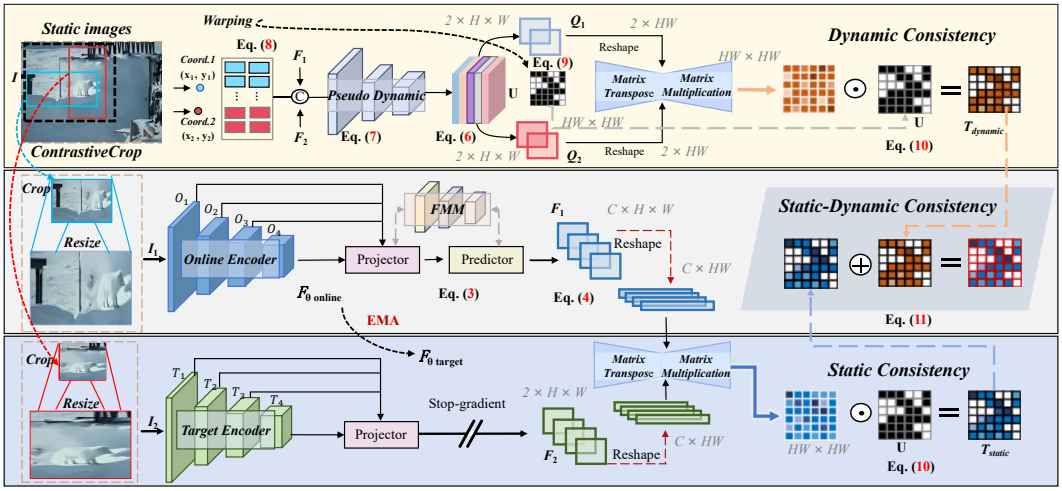


Figure 3: Framework of the proposed model. An image $\boldsymbol{I}$ is transformed into two views, $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, through crop-resize transformations. These views are processed by the online and target networks, $\mathcal{F}_{\boldsymbol{\theta}_{\text{online}}}$ and $\mathcal{F}_{\boldsymbol{\theta}_{\text{target}}}$. The online network uses both projector and predictor heads to generate feature maps $\boldsymbol{F}_1$, while the target network uses only the projector head to generate $\boldsymbol{F}_2$. The pseudo-dynamic signal generation module outputs $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$. These are used to compute the static-dynamic consistency loss by combining feature and dynamic similarities with a positive sample mask.

This design reduces the likelihood of repeatedly cropping around the center of the localized bounding box, which often contains redundant features, thereby increasing the diversity of the sampled positive pairs and improving representation robustness.

Dual-encoder architecture. Once the crops $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ are obtained, they are processed by two encoders: the online encoder $\mathcal{F}_{\boldsymbol{\theta}_{\text{online}}}$ and the target encoder $\mathcal{F}_{\boldsymbol{\theta}_{\text{target}}}$. The two encoders share the same architecture but maintain separate parameters. Both are equipped with lightweight projector heads for mapping backbone features into a contrastive embedding space. Each projector head is implemented as a multi-layer perceptron (MLP) consisting of two fully connected layers followed by `ReLU` activation. The online encoder is optimized by standard backpropagation, whereas the target encoder is updated using an exponential moving average (EMA) (Haynes et al., 2012) of the online encoder parameters. Specifically, the target parameters are updated at each iteration as:

$$\boldsymbol{\theta}_{\text{target}} \leftarrow m\,\boldsymbol{\theta}_{\text{target}} + (1-m)\,\boldsymbol{\theta}_{\text{online}}, \tag{3}$$

where $m \in [0,1]$ denotes the momentum coefficient. In our implementation, m starts from 0.99 and gradually increases to 1 during training. This momentum update allows the target encoder to evolve smoothly and provides stable representations for contrastive learning.

Feature normalization. The online and target branches produce dense feature maps $\boldsymbol{Z}_1 = \mathcal{F}_{\boldsymbol{\theta}_{\text{online}}}(\boldsymbol{I}_1)$ and $\boldsymbol{Z}_2 = \mathcal{F}_{\boldsymbol{\theta}_{\text{target}}}(\boldsymbol{I}_2)$, where $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^{C\times H\times W}$. We apply ℓ_2 normalization along the channel dimension at each spatial location (Celebi et al., 2011):

$$\boldsymbol{F}_1(i) = \frac{\boldsymbol{Z}_1(i)}{\|\boldsymbol{Z}_1(i)\|_2}, \qquad \boldsymbol{F}_2(j) = \frac{\boldsymbol{Z}_2(j)}{\|\boldsymbol{Z}_2(j)\|_2}. \tag{4}$$

where $i, j \in \{1, \ldots, HW\}$ index the flattened spatial locations in the two views. Thus, $\boldsymbol{F}_1, \boldsymbol{F}_2 \in \mathbb{R}^{C\times H\times W}$, and each nonzero local feature vector is normalized to unit length. This removes feature-magnitude differences from the subsequent cosine similarity comparisons.

Contrastive objective. For representation learning, we encourage the embeddings of positive pairs $(\boldsymbol{F}_1, \boldsymbol{F}_2)$ to be aligned. Following recent work (Caron et al., 2021), we omit explicit negative pairs, and use the momentum-updated target encoder to provide a smoothly evolving learning target. Accordingly, we adopt a contrastive objective to enforce cross-view feature consistency:

$$\mathcal{L} = -\frac{\langle \boldsymbol{F}_1, \boldsymbol{F}_2 \rangle}{\|\boldsymbol{F}_1\|_2 \|\boldsymbol{F}_2\|_2}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Since $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$ are ℓ_2-normalized beforehand, Eq. equation 5 is equivalent to maximizing their cosine similarity. This contrastive loss encourages representations from different cropped views of the same image to remain consistent without explicitly constructing negative pairs. Equation equation 5 describes the cosine-alignment objective at a schematic level. Cropping and resizing mean that identical array indices in the two views need not correspond to the same original-image location. We therefore define spatially valid cross-view pairs explicitly using the crop coordinates below; the final dense objective is given in Eq. equation 11.

3.1 Static-Dynamic Correspondence Learning

Static images inherently lack realistic dynamic information, which poses a major challenge for VCOD. To address this issue, we seek to simulate motion cues directly from static images, enabling the model to establish correspondences between static and dynamic representations. Inspired by the human visual system, which often performs multi-view reasoning when perceiving scenes, we propose a pseudo-dynamic signal generation strategy based on crop-induced spatial transformations (Ye et al., 2021; Liu et al., 2022). The key idea is to construct spatial variations within a single static image that serve as a proxy for cross-view correspondence learning. These transformations do not recover physical object motion or actual temporal dynamics. In this subsection, we provide a detailed account of how to model static-dynamic correspondences and how these signals are integrated into our learning framework.

3.1.1 Generation of Pseudo-Dynamic Signals.

Existing VCOD methods typically rely on optical flow estimation to capture motion information between video frames. However, optical flow estimation can be unreliable in camouflaged scenes because of weak object-background contrast, repetitive textures, and ambiguous boundaries. Since our training framework does not use video sequences, we instead simulate motion-like variations from spatial transformations between cropped views of the same static image.

Given the normalized feature maps $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, we introduce a lightweight pseudo-dynamic signal generation module $\mathcal{F}_{\text{pseudo}}$ to estimate bidirectional displacement-like representations between the two views. Specifically, the features from the two views are jointly processed by a two-layer convolutional network. The first convolutional layer is followed by batch normalization and a `ReLU` activation, while the second convolutional layer directly produces a two-channel output without an activation function, allowing signed horizontal and vertical displacement responses. The two input orderings produce the forward and backward pseudo-dynamic representations, respectively. These terms indicate the ordering of the feature inputs rather than physically calibrated optical-flow directions. The signals are formulated as:

$$\boldsymbol{Q}_1 = \mathcal{F}_{\text{pseudo}}(\boldsymbol{F}_1, \boldsymbol{F}_2), \quad \boldsymbol{Q}_2 = \mathcal{F}_{\text{pseudo}}(\boldsymbol{F}_2, \boldsymbol{F}_1), \tag{6}$$

$$\mathcal{L} = -\left\{ \underbrace{\frac{\langle \boldsymbol{F}_1, \boldsymbol{F}_2 \rangle}{\|\boldsymbol{F}_1\|_2 \|\boldsymbol{F}_2\|_2}}_{\text{static correspondence}} + \underbrace{\frac{\langle \boldsymbol{Q}_1, \boldsymbol{Q}_2 \rangle}{\|\boldsymbol{Q}_1\|_2 \|\boldsymbol{Q}_2\|_2}}_{\text{pseudo-dynamic correspondence}} \right\}. \tag{7}$$

where $\boldsymbol{Q}_1, \boldsymbol{Q}_2 \in \mathbb{R}^{2\times H\times W}$ represent the forward and backward pseudo-dynamic signals between $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, respectively. Without explicit optical flow computation, SDC utilizes spatial transformations between overlapping regions to simulate motion. Equation equation 7 summarizes the two alignment terms; the spatially masked, weighted objective is specified in Eq. equation 11.

3.1.2 Selection of Positive Regions.

Here, we describe how we construct positive pairs of feature maps ($\boldsymbol{F}_1, \boldsymbol{F}_2$ and $\boldsymbol{Q}_1, \boldsymbol{Q}_2$). The two crops can cover different regions and scales in the original image, so resizing them to the same resolution changes the correspondence between feature locations. We apply ContrastiveCrop to obtain two views and record their original coordinates. These cropped views and their corresponding coordinates are rescaled and transformed. Based on the spatial positions of the two cropped views in the original image, we generate the transformed pixel coordinates as follows:

$$(\boldsymbol{x}_1^o, \boldsymbol{y}_1^o) = \mathcal{W}_1(\boldsymbol{x}_1, \boldsymbol{y}_1), \quad (\boldsymbol{x}_2^o, \boldsymbol{y}_2^o) = \mathcal{W}_2(\boldsymbol{x}_2, \boldsymbol{y}_2), \tag{8}$$

where $\mathcal{W}_1$ and $\mathcal{W}_2$ denote the inverse coordinate mappings associated with the respective crop-resize transformations, and $\boldsymbol{x}_{\cdot}^{o}, \boldsymbol{y}_{\cdot}^{o} \in \mathbb{R}^{H\times W}$ represent the transformed horizontal and vertical pixel coordinates, respectively. Since we know the original coordinates of these views, we map their feature-grid coordinates back to a common original-image coordinate system. Subsequently, we can compute the Euclidean distance $\mathcal{D}$ between the two sets of feature coordinates in the original image space. This distance measures spatial proximity, rather than feature similarity, and is used to select valid correspondence pairs. To define the spatial neighborhood of the local feature space, we set a positive radius r for the corresponding cropped regions (the overlapping regions of the two cropped views). Finally, the positive region selection mask $\boldsymbol{U} \in \mathbb{R}^{HW\times HW}$ for the two cropped views can be represented as:

$$\boldsymbol{U}(i,j) = \begin{cases} \frac{1}{1+\mathcal{D}(i,j)^2}, & \text{if } \mathcal{D}(i,j) \le r, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where i and j index spatial locations in the first and second views, respectively. We set the positive radius r to 0.3, using the same coordinate scale for r and $\mathcal{D}$. To learn accurate visual correspondences from static images, we design an unsupervised signal (i.e., $\boldsymbol{U}$), guided by the known coordinates of the two cropped views. Combining Eq. equation 9, the static and dynamic correspondence methods can be further expressed as:

$$\boldsymbol{T}_{\text{static}}(i,j) = \frac{\langle \boldsymbol{F}_1(i), \boldsymbol{F}_2(j)\rangle}{\|\boldsymbol{F}_1(i)\|_2\|\boldsymbol{F}_2(j)\|_2}\boldsymbol{U}(i,j), \qquad \boldsymbol{T}_{\text{dynamic}}(i,j) = \frac{\langle \boldsymbol{Q}_1(i), \boldsymbol{Q}_2(j)\rangle}{\|\boldsymbol{Q}_1(i)\|_2\|\boldsymbol{Q}_2(j)\|_2}\boldsymbol{U}(i,j). \tag{10}$$

where i and j denote spatial locations in the two cropped views, respectively. $\boldsymbol{T}_{\text{static}} \in \mathbb{R}^{HW\times HW}$ and $\boldsymbol{T}_{\text{dynamic}} \in \mathbb{R}^{HW\times HW}$ represent the masked pairwise similarities of static and pseudo-dynamic representations, respectively. The positive mask $\boldsymbol{U}$ retains only spatially valid correspondence pairs determined by the crop coordinates. Our objective is to maximize the similarities of these valid static and dynamic correspondences.

Loss Function. Learning correspondences from static images and applying them to VCOD tasks is not trivial. To compensate for the absence of real motion in static images, we generate pseudo-dynamic signals from cropped image views and jointly model static and pseudo-dynamic correspondences. Based on the pairwise correspondence matrices defined in Eq. equation 10, we formulate the static-dynamic consistency loss as:

$$\begin{aligned} \mathcal{L}_{\text{static}} &= -\frac{\sum_{i=1}^{HW}\sum_{j=1}^{HW}\boldsymbol{T}_{\text{static}}(i,j)}{\sum_{i=1}^{HW}\sum_{j=1}^{HW}\boldsymbol{U}(i,j)}, \qquad \mathcal{L}_{\text{dynamic}} = -\frac{\sum_{i=1}^{HW}\sum_{j=1}^{HW}\boldsymbol{T}_{\text{dynamic}}(i,j)}{\sum_{i=1}^{HW}\sum_{j=1}^{HW}\boldsymbol{U}(i,j)}, \\ \mathcal{L}_{\text{sdc}} &= \mathcal{L}_{\text{static}} + \beta\mathcal{L}_{\text{dynamic}}. \end{aligned} \tag{11}$$

where $\mathcal{L}_{\text{static}}$ and $\mathcal{L}_{\text{dynamic}}$ maximize the similarities of spatially valid static and pseudo-dynamic correspondence pairs, respectively, as determined by the positive mask $\boldsymbol{U}$. The hyperparameter β controls the contribution of the dynamic correspondence loss $\mathcal{L}_{\text{dynamic}}$ and is empirically set to 1. The final objective $\mathcal{L}_{\text{sdc}}$ jointly optimizes static and pseudo-dynamic correspondence learning, thereby encouraging consistent representations across the two cropped views. The weighted averages in Eq. equation 11 are defined for crop pairs with $\sum_{i,j}\boldsymbol{U}(i,j) > 0$. The online branch is optimized by backpropagation, whereas the target encoder is updated using Eq. equation 3.

4 EXPERIMENTS

Training. As stated in our method, we use the benchmark COD image dataset, rather than video datasets, for training our model, which combines 3,040 images from COD10K (Fan et al., 2020a) and 1,000 images from CAMO (Le et al., 2019) as the static training dataset. During training, label-free raw images are randomly sampled. We create two randomly generated crops using spatial augmentation and resize all crops to 256×256. The model is trained for 100 epochs with a batch size of 16. The Adam optimizer is employed with an initial learning rate of 10^{-3}. During training, the online encoder is updated with gradients, while the parameters of the target encoder are adjusted through an EMA with a coefficient starting at 0.99 and gradually increasing to 1.

Table 1: Quantitative comparison with state-of-the-art methods on MoCA–Mask and CAD2016. ↑ (↓) indicates that higher (lower) is better. **Bold** denotes the best result among unsupervised methods.

Method	Pub.	Sup.	Input	Param.	FPS	MoCA–Mask						CAD2016					
						S_α	E_ϕ	F_β^w	M	mDice	mIoU	S_α	E_ϕ	F_β^w	M	mDice	mIoU
SINet (Fan et al., 2020b)	CVPR'20	Fully	352^2	49M	6	.574	.655	.185	.030	.221	.156	.601	.589	.204	.089	.289	.209
SINet-v2 (Fan et al., 2021)	TPAMI'22	Fully	352^2	27M	37	.571	.608	.175	.035	.211	.153	.544	.546	.181	.049	.170	.110
SLT-Net (Cheng et al., 2022b)	CVPR'22	Fully	352^2	82M	36	.656	.785	.357	.021	.387	.310	.679	.805	.420	.033	.445	.342
IMEX (Hui et al., 2024a)	TMM'24	Fully	352^2	32M	180	.661	.778	.371	.020	.409	.319	.684	.813	.452	.033	.469	.370
TSP-SAM (Hui et al., 2024b)	CVPR'24	Fully	352^2	–	3	.689	.808	.444	.008	.458	.388	.704	.912	.524	.028	.543	.438
ZoomNeXt (Pang et al., 2024)	TPAMI'24	Fully	384^2	84M	26	.734	.736	.476	.010	.497	.422	.757	.865	.593	.020	.599	.510
CamSAM2 (Zhou et al., 2025)	NeurIPS'25	Prompt	1024^2	39.4M	–	.829	.942	.724	.006	.755	.648	.872	.963	.795	.013	.796	.692
EventVCOD (Zhang et al., 2026)	AAAI'26	Fully	1024^2	98.10M	2.56	.753	.855	.573	.009	.574	.496	.802	.887	.717	.023	.717	.615
SLT-Net-ST (Cheng et al., 2022c)	CVPR'22	Weakly	352^2	82M	36	.636	.705	.303	.026	.372	.284	.694	.845	.497	.030	.484	.392
RCRNet (Yan et al., 2019)	ICCV'19	Semi	448^2	54M	6	.597	.583	.174	.025	.194	.137	.596	.582	.172	.026	.309	.229
EMIP (Zhang et al., 2025b)	TIP'25	Semi	352^2	–	–	.669	.667	.374	.017	.424	.326	.710	.846	.504	.029	.528	.415
MG (Yang et al., 2021)	ICCV'21	Unsup.	128×224	5M	6	.547	.537	.165	.095	.197	.141	.484	.558	.314	.370	.351	.260
Ours (SDC)	–	**Unsup.**	256^2	48M	339	**.671**	**.706**	**.444**	**.008**	**.435**	**.341**	**.721**	**.873**	**.604**	**.018**	**.584**	**.479**

Testing. Following previous methods (Hui et al., 2024b; Cheng et al., 2022a), we have conducted a evaluation on two widely used video datasets: MoCA-Mask (Cheng et al., 2022a) and CAD2016 (Bideau & Learned-Miller, 2016). The MoCA-Mask dataset is a video camouflaged object detection dataset, covering camouflaged animals moving in natural scenes. It contains 87 video sequences (22,939 frames), of which 71 video sequences are used for training and 16 sequences for testing. It is well annotated with bounding boxes and pixel-level segmentation masks on every fifth frame. Meanwhile, the Camouflaged Animal Dataset (CAD2016) is a small video camouflaged object detection dataset extracted from YouTube videos. It consists of nine short video clips in total and is well annotated with pixel-level semantic masks on every fifth frame. Notably, these datasets are only used for testing, and the training sets are not utilized in our method.

Architecture We use ResNet-50 (He et al., 2016) as the backbone for feature extraction, removing the strides in the last two residual blocks (res3 and res4) to enhance feature resolution. The feature mapping module then constructs the projector and predictor heads to generate feature embeddings from the backbone output. To ensure consistent feature representation, we apply Euclidean normalization to the obtained embeddings.

Evaluation Metrics. During the testing phase, we evaluate the predicted masks using six widely adopted metrics: S-measure (S_α) (Fan et al., 2017), Weighted F-measure (F_β^w) (Margolin et al., 2014), Enhanced-alignment measure (E_ϕ) (Achanta et al., 2009), Mean Absolute Error (MAE, M) (Perazzi et al., 2012), mean Dice (mDice), and mean Intersection over Union (mIoU). A better VCOD method yields higher scores in S_α, F_β^w, E_ϕ, mDice, and mIoU, while producing a lower MAE score.

4.1 Comparison With the State-of-the-Arts

Quantitative Results. For a fair comparison, all results are taken from the original papers or obtained using the released models under the same evaluation protocol, without post-processing. As shown in Table 1, SDC is trained solely on unlabeled static images, requiring neither video training data nor manual annotations. Despite this challenging setting, SDC outperforms all compared weakly supervised, semi-supervised, and unsupervised VCOD methods, while remaining competitive with or surpassing most representative fully supervised methods. In addition, SDC requires only about one hour of training on a single RTX 4090 GPU and achieves the highest inference speed of 339 FPS with relatively few parameters, demonstrating an excellent balance among accuracy, efficiency, and annotation-free training.

Qualitative Results. To further validate our method qualitatively, we provide visual comparisons with representative methods in Fig. 4. Our method not only effectively highlights camouflaged objects but also delivers more accurate boundary segmentation. Particularly in regions with weak texture contrast or partial occlusion, our results remain structurally coherent and visually clear, outperforming both fully and weakly supervised baselines.

4.2 Ablation Study

We conduct ablation studies on the MoCA–Mask dataset with $T{=}5$ to investigate the contributions of the proposed components, training strategies, and major hyperparameters.

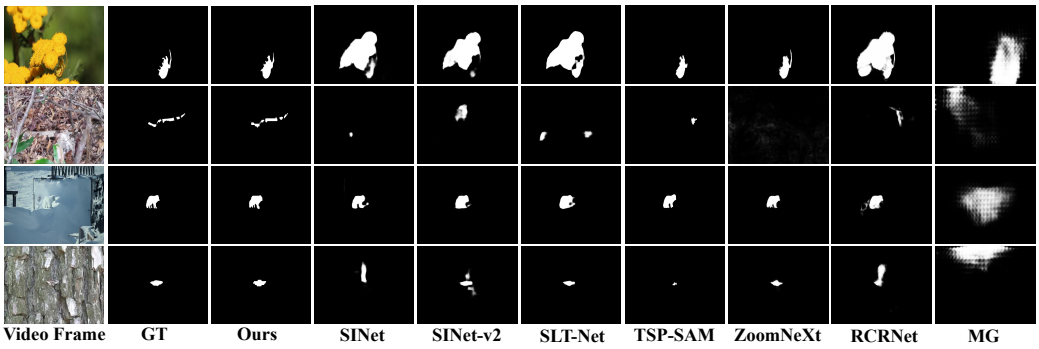


Figure 4: Qualitative comparisons of our proposed method and 7 state-of-the-art VCOD methods across representative scenarios.

Table 2: Ablation studies on MoCA–Mask. B, S, D, and V denote the baseline, static correspondence, dynamic correspondence, and video training data, respectively. The default configuration in each study is highlighted in light gray.

B	S	D	S_α	E_ϕ	F_β^w	M
✓			.661	.761	.381	.025
✓	✓		.685	.776	.432	.016
✓	✓	✓	**.724**	**.811**	**.495**	**.008**

(a) Critical modules.

Augmentation	S_α	E_ϕ	F_β^w	M
Cont. Crop	**.724**	**.811**	**.495**	**.008**
Transform.	.712	.794	.463	.011
Color Jitter	.705	.781	.454	.013
Gauss. Blur	.686	.773	.442	.015

(b) Augmentation strategies.

β	S_α	E_ϕ	F_β^w	M
0.1	.681	.760	.432	.017
0.5	.709	.797	.459	.010
1.0	**.724**	**.811**	**.495**	**.008**
2.0	.674	.753	.424	.018

(c) Dynamic-loss weight β.

r_p	S_α	E_ϕ	F_β^w	M
.01	.701	.774	.470	.013
.10	.710	.783	.482	.010
.30	**.724**	**.811**	**.495**	**.008**
.50	.705	.777	.471	.011

(d) Positive radius r_p.

Data	V	S_α	E_ϕ	F_β^w	M
MoCA–Mask	✓	.721	.807	.491	.009
COD (static)	×	**.724**	**.811**	**.495**	**.008**

(e) Training data.

Dim.	S_α	E_ϕ	F_β^w	M
128	.711	.803	.479	.009
256	**.724**	**.811**	.495	**.008**
512	.716	.806	**.502**	**.008**

(f) Hidden dimension of FMM.

1) Analysis of Critical Modules. As shown in Table 2(a) and Fig. 5, introducing static correspondence consistently improves the baseline, while further incorporating dynamic correspondence yields additional performance gains. These results demonstrate that static and dynamic correspondence provide complementary cues for learning effective unsupervised representations.

2) Effect of Augmentation Strategies. Table 2(b) compares different augmentation strategies. ContrastiveCrop achieves the best overall performance, as overlapping views preserve appearance consistency and provide reliable spatial correspondences. In contrast, color jittering and Gaussian blur disturb camouflage-related appearance cues and degrade performance.

3) Dynamic Loss. As shown in Table 2(c), the model achieves the best performance when the dynamic-loss weight is set to β=1.0. A smaller weight weakens dynamic correspondence learning, whereas an excessively large weight introduces noisy pseudo-dynamic cues. These results indicate the importance of balancing the static and dynamic objectives.

4) Positive Discriminant. Table 2(d) studies the positive radius r_p used for correspondence discrimination. A small radius provides insufficient positive pairs, while a large radius introduces incorrect matches. The best performance is obtained at r_p=0.30, providing an appropriate balance for correspondence learning.

5) Training Dataset. Table 2(e) compares training on MoCA–Mask videos and static COD images. Training solely on static COD images achieves better overall performance, demonstrating that SDC can learn effective VCOD representations without relying on video training data.

6) Feature Mapping. Table 2(f) evaluates different hidden dimensions of the Feature Mapping Module. A hidden dimension of 256 provides the best overall trade-off between segmentation performance and computational cost. A smaller dimension limits representation capacity, whereas a larger dimension introduces additional overhead with only marginal performance benefits.

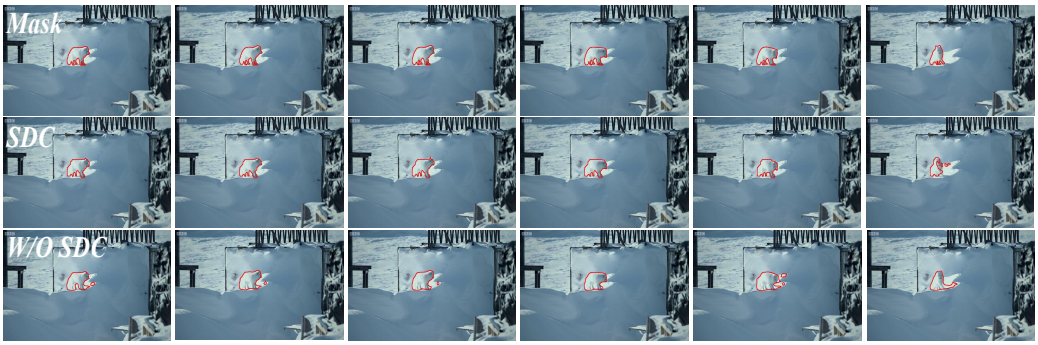


Figure 5: Ablation study quantitatively comparing the effect of SDC on video frames using the MoCA–Mask dataset with T=5.

Table 3: End-point error (EPE) between the pseudo-dynamic fields and reference fields generated by RAFT and PWCNet on MoCA–Mask and CAD2016. Lower values indicate better consistency.

Model	**Direction**	**EPE ↓**	
		MoCA–Mask	**CAD2016**
ΔSDC–RAFT (Teed & Deng, 2020)	Forward	6.23	13.38
ΔSDC–RAFT (Teed & Deng, 2020)	Backward	5.34	12.87
ΔSDC–PWCNet (Sun et al., 2018)	Forward	4.58	6.65
ΔSDC–PWCNet (Sun et al., 2018)	**Backward**	**4.19**	**6.53**

4.3 Pseudo-Dynamic Signals

Although SDC is trained solely on static images, it introduces pseudo-dynamic signals to learn displacement-aware correspondences. The generator $\mathcal{F}_{\text{pseudo}}$ contains two convolutional layers, each followed by batch normalization and ReLU. Given two cropped views of the same image, it predicts a dense correspondence field without video data, flow annotations, or an additional optical-flow network during training. To examine whether the pseudo-dynamic signals contain meaningful displacement information, we use RAFT and PWCNet only as offline reference models. For each individual frame from the MoCA–Mask and CAD2016 validation sets, two cropped views are generated following the same augmentation strategy used during training. RAFT and PWCNet are then applied to these paired views to produce reference correspondence fields. Neither temporally adjacent frames nor optical-flow predictions are used to train SDC. As shown in Table 3, the pseudo-dynamic fields maintain consistent agreement with both reference models in the forward and backward directions. These results demonstrate that $\mathcal{F}_{\text{pseudo}}$ captures meaningful displacement patterns from paired views of a single static image, enabling dynamic-aware correspondence learning under a static-image-only training setting. Additional visualizations of the pseudo-dynamic fields and qualitative VCOD results are provided in the supplementary material.

5 Conclusion

In this paper, we introduced SDC, an efficient and unsupervised framework for VCOD. Unlike previous methods that rely on large-scale labeled datasets or complex video data processing techniques, SDC effectively learns from unlabeled static images. By integrating static and dynamic visual correspondence learning, our approach simulates dynamic signals from static images, ensuring static-dynamic consistency throughout the learning process. Our method leverages pseudo-dynamic learning strategies to capture the temporal consistency of objects, even when working with still images. Through extensive experiments on public benchmarks, SDC outperforms existing video-dependent methods, demonstrating its ability to detect camouflaged objects in videos effectively using only static images during training. In conclusion, SDC provides a robust, scalable, and annotation-free method to VCOD, illustrating the potential of static images for dynamic video tasks and paving the way for further advancements in unsupervised VCOD.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language translation, improve the readability and clarity of the manuscript, and provide limited feedback on research methodology and experimental design. AI-generated suggestions were used only as references and did not replace the authors' independent research decisions. All AI-assisted translations and revisions were manually reviewed for technical accuracy, terminology, and consistency, and all methodological and experimental decisions were made and verified by the authors. Generative AI was not used to generate experimental results, formulate mathematical claims, implement the proposed method, conduct data analysis, or interpret the reported results. The authors take full responsibility for the final content, claims, and conclusions of this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work. The experimental section provides details of the datasets, evaluation metrics, model configurations, training settings, and implementation protocols used in our experiments. Additional implementation details, parameter analyses, and supplementary experiments are provided in the Appendix. The anonymized source code is also included in the supplementary material for double-blind review. Together, these materials are intended to facilitate independent reproduction and verification of the reported results.

REFERENCES

R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pp. 1597–1604, 2009.

P. Bideau and E. Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *Proc. Eur. Conf. Comput. Vis.*, pp. 433–449, 2016.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. pp. 9650–9660, 2021.

M. Emre Celebi, Fatih Celiker, and Hassan A. Kingravi. On euclidean norm approximations. *Pattern Recognition*, 44(2):278–283, 2011. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2010.08.028.

Huafeng Chen, Chenguang Zhu, Yueming Lyu, and Caifeng Shan. Beyond weak supervision: Mllms-guided graded knowledge distillation for unsupervised camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27547–27557, 2026.

Tianyou Chen, Hui Ruan, Shaojie Wang, Jin Xiao, and Xiaoguang Hu. A three-stage model for camouflaged object detection. *Neurocomputing*, 614:128784, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2024.128784.

X. Cheng et al. Implicit motion handling for video camouflaged object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 13864–13873, 2022a.

Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13864–13873, 2022b.

Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13854–13863, 2022c. doi: 10.1109/CVPR52688.2022.01349.

Ji Du, Fangwei Hao, Mingyang Yu, Desheng Kong, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Shift the lens: Environment-aware unsupervised camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19271–19282, 2025.

D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2774–2784, 2020a.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2777–2787, 2020b.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(10):6024–6042, 2021.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022a.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022b. doi: 10.1109/TPAMI.2021.3085766.

David Haynes, Steven Corns, and Ganesh Kumar Venayagamoorthy. An exponential moving average algorithm. In *2012 IEEE Congress on Evolutionary Computation*, pp. 1–8, 2012. doi: 10.1109/CEC.2012.6252962.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Wenjun Hui, Zhenfeng Zhu, Guanghua Gu, Meiqin Liu, and Yao Zhao. Implicit-explicit motion learning for video camouflaged object detection. *IEEE Transactions on Multimedia*, 26:7188–7196, 2024a. doi: 10.1109/TMM.2024.3361170.

Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, and Yao Zhao. Endow sam with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19058–19067, 2024b. doi: 10.1109/CVPR52733.2024.01803.

Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. volume 33, pp. 19545–19560, 2020.

Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI, pp. 142–152, 2021.

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*, 184:45–56, 2019. ISSN 1077-3142.

Shuaicheng Liu, Nianjin Ye, Chuan Wang, Jirong Zhang, Lanpeng Jia, Kunming Luo, Jue Wang, and Jian Sun. Content-aware unsupervised deep homography estimation and its extensions. 45 (3):2849–2863, 2022.

Zelin Lu, Liang Xie, Xing Zhao, Binwei Xu, Haoran Liang, and Ronghua Liang. A weakly-supervised cross-domain query framework for video camouflage object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1506–1518, 2025.

R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *CVPR*, 2014.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2160–2170, 2022.

Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024. doi: 10.1109/TPAMI.2024.3417329.

Xiangyu Peng, Kai Wang, Zheng Zhu, Mang Wang, and Yang You. Crafting better contrastive views for siamese representation learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16010–16019, 2022. doi: 10.1109/CVPR52688.2022.01556.

F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pp. 733–740, 2012.

Adam Steel, Madeleine M. Billings, Edward H. Silson, and Caroline E. Robertson. A network linking scene perception and spatial memory systems in posterior cerebral cortex. *Nature Communications*, 12(1):2632, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22848-z.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943, 2018. doi: 10.1109/CVPR.2018.00931.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 402–419, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

Haoke Xiao, Lv Tang, Bo Li, Zhiming Luo, and Shaozi Li. Disentangled self-supervised video camouflaged object detection and salient object detection. *Neural Networks*, 194:108077, 2026. doi: 10.1016/j.neunet.2025.108077.

Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7283–7292, 2019. doi: 10.1109/ICCV.2019.00738.

Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7157–7168, 2021. doi: 10.1109/ICCV48922.2021.00709.

Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. pp. 13117–13125, 2021.

Hong Zhang, Yixuan Lyu, Hanyang Liu, Jianbo Song, Ding Yuan, and Yifan Yang. Towards explainable video camouflaged object detection: Sam2 with eventstream-inspired data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12511–12519, 2026. doi: 10.1609/aaai.v40i15.38245.

Miao Zhang, Beiqi Hu, Shunyu Yao, Yongri Piao, and Huchuan Lu. Pmnet: Predator-mimicking network for video camouflaged object detection. *IEEE Transactions on Multimedia*, 27:2374–2383, 2025a. doi: 10.1109/TMM.2024.3521775.

Xin Zhang, Tao Xiao, Ge-Peng Ji, Xuan Wu, Keren Fu, and Qijun Zhao. Explicit motion handling and interactive prompting for video camouflaged object detection. *IEEE Transactions on Image Processing*, 34:2853–2866, 2025b.

Junmin Zhong, Anzhi Wang, Chunhong Ren, and Jintao Wu. A survey on deep learning-based camouflaged object detection. *Multimedia Syst.*, 30(5), September 2024. ISSN 0942-4962.

Yuli Zhou, Yawei Li, Yuqian Fu, Luca Benini, Ender Konukoglu, and Guolei Sun. Camsam2: Segment anything accurately in camouflaged videos. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1855.

A EXPERIMENTS

A.1 DATASET

A.1.1 TRAINING DATASET.

We use a static image training dataset consisting of 3,040 images from COD10K (Fan et al., 2020a) and 1,000 images from CAMO (Le et al., 2019), which are utilized for training our model. It is important to note that we do not use any annotations or ground truth information, during training. This means that, although the datasets themselves provide pixel-level annotations, our approach does not rely on these annotations for training.

- **CAMO:** The CAMO dataset (Le et al., 2019) is designed specifically for camouflaged object segmentation tasks. It consists of 1,250 images, with 1,000 images for training and 250 images for testing. The dataset focuses on two types of camouflaged objects: natural camouflaged objects, which typically correspond to animals in the real world, and man-made camouflaged objects, often corresponding to humans. The images are annotated with pixel-level masks for each camouflaged object.
- **COD10K:** The COD10K dataset (Fan et al., 2020a) is the largest dataset for camouflaged object detection to date, containing 10,000 images, with 6,000 images for training and 4,000 images for testing. The dataset covers a wide range of camouflaged objects across different scenes and includes 78 categories of camouflaged objects. The images are sourced from various photography websites, ensuring diverse and realistic representations.

A.1.2 TESTING DATASETS.

For testing, we evaluate our method on two widely used video datasets: MoCA-Mask (Cheng et al., 2022a) and CAD2016 (Bideau & Learned-Miller, 2016).

- **MoCA-Mask:** A video camouflaged object detection dataset consisting of 87 video sequences (22,939 frames), where camouflaged animals move in natural scenes. The dataset is annotated with bounding boxes and pixel-level segmentation masks on every fifth frame. 71 video sequences are used for training, while 16 sequences are reserved for testing.
- **CAD2016:** A smaller video camouflaged object detection dataset, extracted from YouTube videos, containing nine short video clips. It is annotated with pixel-level semantic masks on every fifth frame.

Notably, these video datasets are only used for testing, and the training sets are not utilized in our method.

A.2 TRAINING PROCEDURE

Algorithm 1 summarizes the PyTorch-like training procedure of SDC. Given a static image, the data loader generates two cropped views together with their spatial coordinates. The online and target encoders extract normalized dense features from both views, while the target branch is updated by exponential moving average and does not receive gradient updates.

For each pair of cross-view features, SDC first constructs a distance-aware positive mask according to the crop coordinates and the positive radius r. The pseudo-dynamic module $\mathcal{F}_{\mathrm{pseudo}}$ then predicts bidirectional displacement-aware representations between the two views. Static correspondence is measured directly from the encoder features, whereas dynamic correspondence is computed from the generated pseudo-dynamic representations. Their similarities are restricted by the positive mask and jointly optimized through the static-dynamic consistency objective.

The loss is computed symmetrically in both view directions. During optimization, gradients update only the online encoder and the pseudo-dynamic module, while the target encoder is updated using the momentum coefficient m. Therefore, the entire framework learns motion-aware correspondence from paired crops of static images, without requiring video data, optical-flow supervision, or an additional pre-trained flow estimator. We provide the PyTorch-like pseudocode in Algorithm 1.

A.3 QUALITATIVE RESULTS.

To further demonstrate the effectiveness of our SDC framework in VCOD, we present qualitative results comparing our method with the ground truth annotations (Mask) from the MoCA-Mask dataset. As shown in Fig. 6, we compare the results generated by our approach (Ours) with the ground truth masks (Mask). The visual examples highlight how our method excels in detecting camouflaged objects, offering improved segmentation accuracy in complex video sequences.

Algorithm 1 Static-Dynamic Correspondence Network (SDC) PyTorch-like pseudocode.

```
# fo, ft: online and target encoder networks
# r: positive radius
# m: momentum coefficient
# tau: temperature
# pseudo: pseudo-dynamic module
# loader: provides two cropped views and coordinates

for v1, v2, c1, c2 in loader:
    # online features (normalized)
    f1, f2 = normalize(fo(v1)), normalize(fo(v2))
    # target features (normalized)
    t1, t2 = normalize(ft(v1)), normalize(ft(v2))

    # compute static-dynamic consistency loss
    loss = static_dynamic_consistency_loss(f1, t2, c1, c2, r, tau) + \
           static_dynamic_consistency_loss(f2, t1, c2, c1, r, tau)

    # back-propagate
    loss.backward()
    update(fo)                          # update online encoder
    ft = m * ft + (1.0 - m) * fo        # EMA update target encoder

def static_dynamic_consistency_loss(f, t, c1, c2, r, tau):
    # ---- Step 1: generate pseudo-dynamic signals (Eq.3) ----
    q1, q2 = pseudo(f, t), pseudo(t, f)

    # ---- Step 2: construct positive region mask (Eq.4) ----
    D = dist(c1, c2)                      # Euclidean distance
    U = ((D <= r).float() / (1 + D**2)).detach()

    # ---- Step 3: compute static & dynamic similarities (Eq.5) ----
    static  = log_sim(f, t, tau) * U
    dynamic = log_sim(q1, q2, tau) * U

    # ---- Step 4: static-dynamic consistency loss ----
    loss = - (static.mean() + dynamic.mean())
    return loss
```

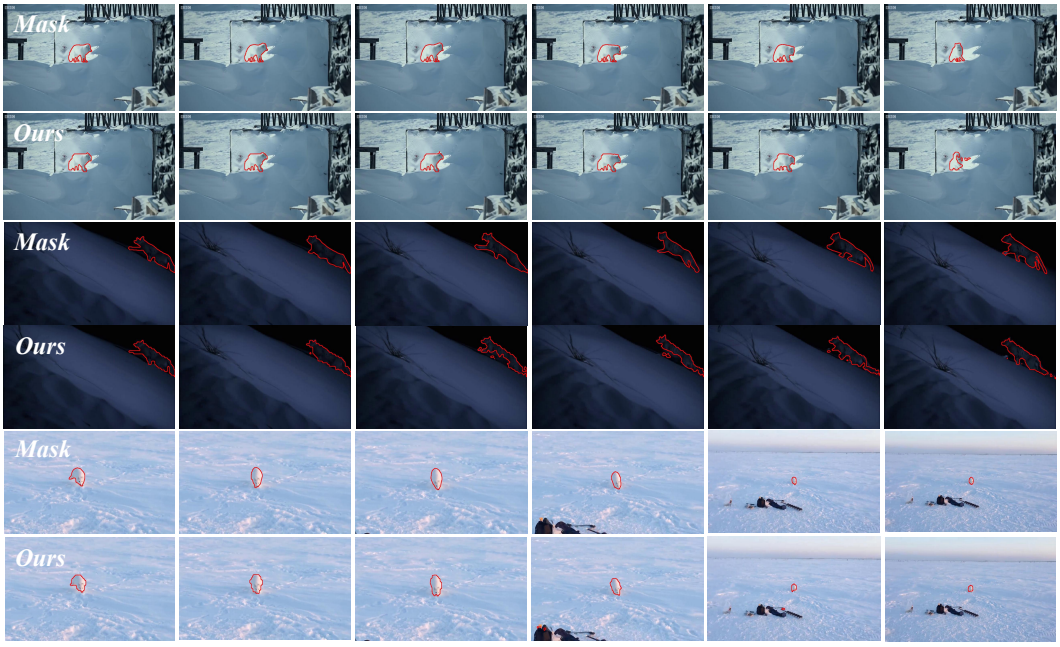


Figure 6: Additional qualitative results for video camouflaged object detection.