arXiv:2507.19789v1 [cs.CV] 26 Jul 2025

TransFlow: Motion Knowledge Transfer from Video Diffusion Models to Video Salient Object Detection

Suhwan Cho[1] Minhyeok Lee[2] Jungho Lee[2] Sunghun Yang[2] Sangyoun Lee[2]

[1] GenGenAI [2] Yonsei University

https://github.com/suhwan-cho/TransFlow

Abstract

Video salient object detection (SOD) relies on motion cues to distinguish salient objects from backgrounds, but training such models is limited by scarce video datasets compared to abundant image datasets. Existing approaches that use spatial transformations to create video sequences from static images fail for motion-guided tasks, as these transformations produce unrealistic optical flows that lack semantic understanding of motion. We present TransFlow, which transfers motion knowledge from pre-trained video diffusion models to generate realistic training data for video SOD. Video diffusion models have learned rich semantic motion priors from large-scale video data, understanding how different objects naturally move in real scenes. TransFlow leverages this knowledge to generate semantically-aware optical flows from static images, where objects exhibit natural motion patterns while preserving spatial boundaries and temporal coherence. Our method achieves improved performance across multiple benchmarks, demonstrating effective motion knowledge transfer.

1. Introduction

The success of deep learning in video processing relies heavily on large-scale datasets. While image datasets are abundant and relatively easy to annotate, collecting and labeling video data remains costly and labor-intensive. To address this challenge, researchers often simulate video sequences from image datasets using spatial transformations such as affine and spline warping. However, these methods fail for motion-guided tasks like video salient object detection (SOD), where realistic optical flow is essential for distinguishing salient objects from backgrounds.

Flow-guided video SOD models require optical flow maps to capture motion dynamics, but existing image datasets lack this crucial temporal information. While recent video segmentation datasets [14, 15, 37] that have advanced the video understanding field can provide optical flow, their segmentation masks are misaligned with salient object annotations, as they focus on general object segmentation rather than salient object segmentation. This fundamental mismatch between available annotations and SOD requirements leaves the data scarcity challenge for video SOD largely unaddressed, consequently constraining model training and hindering both performance and generalization capabilities across diverse video scenarios.

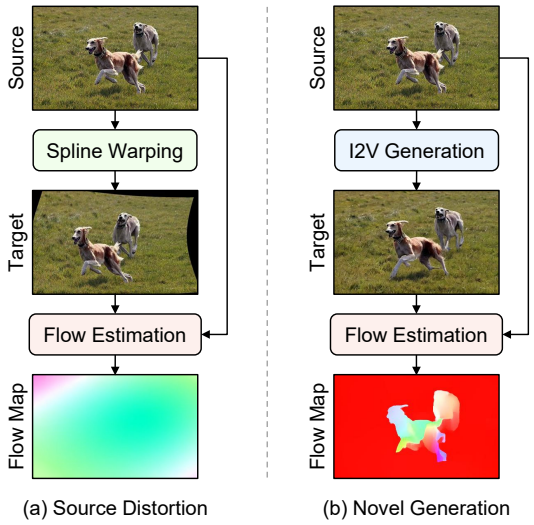


Figure 1. Motion generation comparison: (a) geometric transformations; (b) TransFlow's semantic motion generation.

The fundamental limitation of current flow simulation lies in its geometric approach to motion generation. Spatial transformations like affine and spline warping enforce uniform motion patterns across entire images, treating them as rigid entities rather than understanding that different objects should move independently according to their semantic properties. This produces unrealistic optical flows that lack meaningful motion cues, as illustrated in Figure 1 (a).

To address this limitation, we propose TransFlow, a framework that transfers motion knowledge from pre-trained video diffusion models to generate realistic train-

ing data for video SOD. Unlike geometric transformations, video diffusion models have learned semantic motion understanding from large-scale video data, capturing how different objects naturally move in real scenes. TransFlow leverages this knowledge to generate semantically-aware optical flows from static images, where objects exhibit contextually-appropriate motion patterns while preserving spatial boundaries, as shown in Figure 1 (b). This represents a paradigm shift from geometric to semantic motion generation, enabling the creation of training datasets with meaningful motion cues.

We create a large-scale dataset of image-flow-mask triplets derived from existing image SOD datasets using TransFlow. This allows flow-guided video SOD models to benefit from abundant image data while obtaining realistic motion supervision. Experimental results show improved performance on multiple benchmark datasets, suggesting that transferring motion knowledge from generative models can effectively address data limitations in motion-guided video understanding tasks.

Our main contributions are as follows:

- We propose TransFlow, a framework for transferring semantic motion knowledge from video diffusion models to enhance video SOD training data.
- We demonstrate that video diffusion models can generate semantically-aware motion patterns from static images, producing realistic optical flows that capture independent object motion.
- We explore cross-paradigm knowledge transfer from generative to discriminative video understanding, showing how motion knowledge can be effectively harvested from foundation models.

2. Related Work

Utilizing image data for video models. To leverage the extensive knowledge embedded in large-scale image datasets, video processing tasks commonly incorporate image data through spatial transformations, treating original and transformed images as video sequences.

In video object segmentation (VOS), MaskTrack [46] applies affine transformations and thin-plate spline [4] deformations to segmentation masks, using transformed masks as previous frame annotations alongside query frame images for training. RGMP [41] extends this approach by overlaying foreground objects from salient object detection datasets [9, 51] onto PASCAL-VOC [16, 23] backgrounds while applying similar transformations. Recent methods [7, 8, 10, 39, 42, 49, 50] generate video-like samples by independently transforming images and masks, then sequencing them temporally, demonstrating effectiveness across both semi- and unsupervised VOS settings.

Beyond VOS, STEm-Seg [2] synthesizes training clips for video instance segmentation using random affine transformations and motion blur effects, while FCNS [59] employs similar augmentation techniques to enhance saliency learning and reduce overfitting.

However, these geometric transformation approaches treat images as rigid entities, applying uniform motion patterns without considering semantic properties of different objects. Consequently, the resulting optical flows lack meaningful motion cues that reflect how objects naturally move in real scenes, limiting their effectiveness for motion-guided tasks.

Motion-guided architectures. Motion cues provide crucial insights into object dynamics, typically captured through optical flow estimation between consecutive frames. This temporal information is particularly valuable for segmentation tasks, where flow maps preserve spatial details that improve pixel-level prediction accuracy.

Recent methods integrate motion and appearance information through various fusion mechanisms. MATNet [68] transforms appearance features into motion-attentive representations using asymmetric attention blocks, while MGA [36] employs separate processing modules for each modality and adapts them through end-to-end training. AMC-Net [63] dynamically balances appearance and motion cues via gating functions, and FSNet [29] enforces mutual constraints between multi-modal features during encoding. Methods such as PMN [34], GSA-Net [35], and DPA [12] further leverage attention mechanisms to fuse cross-modal information while capturing temporal dependencies across video sequences.

Despite these architectural advances, flow-guided methods remain constrained by limited training data with realistic motion patterns. Traditional image-to-video simulation produces optical flows from geometric transformations that lack authentic motion characteristics, reducing their effectiveness for training robust motion-guided models.

Image-to-video generation. Video generation models have gained significant attention for content creation and editing applications. Image-to-video generation models, which produce video sequences from single images, demonstrate superior quality and controllability by using input images as explicit first-frame guides.

Diffusion models [26, 47, 52] have significantly advanced image-to-video generation capabilities. Representative approaches include Stable Video Diffusion [3], Gen-2 [48], I2VGen-XL [65], PikaLabs [33], SparseCtrl [22], and SORA [5]. Some methods extend controllability through explicit optical flow or trajectory guidance [38, 64].

Video diffusion models have learned rich semantic motion priors from large-scale video data, encoding understanding of how different objects naturally move in diverse scenes. However, existing work primarily focuses on content generation rather than leveraging this embedded motion knowledge for discriminative video understanding tasks.

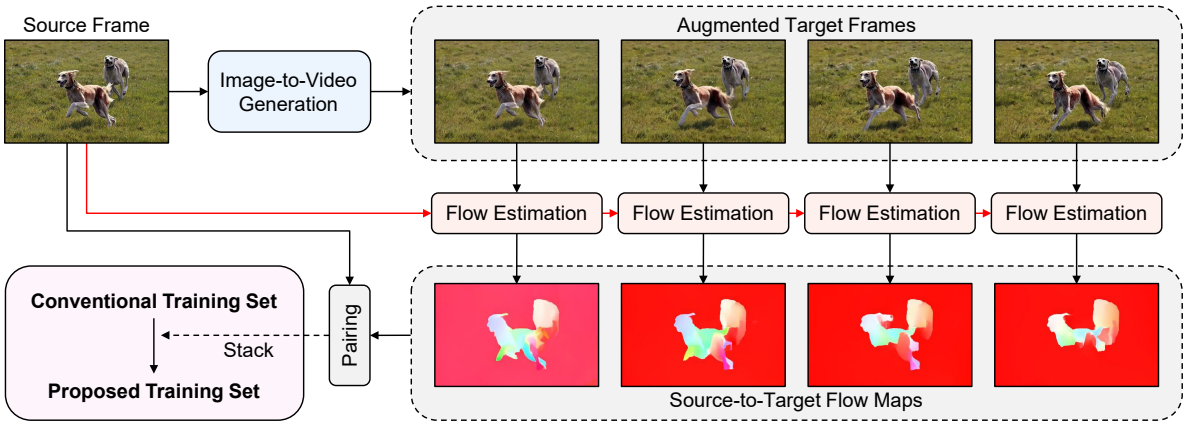


Figure 2. Training data generation pipeline. Static images are transformed into video sequences using diffusion models, then optical flows are estimated between source and generated frames to create image-flow training pairs.

Our work explores transferring this motion understanding to generate realistic training data for video SOD.

3. Approach

3.1. Generating Target Frames

As shown in Figure 1, spatially distorting the original source image to generate a target frame for optical flow estimation does not yield realistic optical flow maps. To obtain high-quality optical flow maps with video-like properties from static images, we use an image-to-video generation model that generates new frames directly, rather than relying on distortion of the source image.

Network overview. In the image-to-video generation process, we employ Stable Video Diffusion [3] based on a 3D-UNet [13] architecture. Following recent approaches, pixel-to-latent conversion is applied before the denoising model, with latent-to-pixel conversion following it, using a pre-trained VAE [31] encoder and decoder, respectively. The VAE decoder incorporates 3D convolutional layers to account for the temporal dimension, ensuring temporal consistency during the decoding process. Starting from Gaussian noise $x \in \mathbb{R}^{C\times T\times H\times W}$ and the source image latent $z_s \in \mathbb{R}^{C\times H\times W}$, the denoised latent $z \in \mathbb{R}^{C\times T\times H\times W}$ is obtained as:

$$z = \Phi(x, z_s^{(T)}), \tag{1}$$

where Φ denotes the denoising 3D-UNet model and $z_s^{(T)}$ represents the temporal replication of z_s across T frames. The output denoised latent z is then decoded with the VAE decoder on a per-frame basis, with each frame representing a transformed image generated from the source image. The number of video frames T is set to 14, following the default setting in Stable Video Diffusion.

Sampling details. During the sampling process, we use 25 steps of the deterministic DDIM sampler [53]. The classifier guidance scale [25] is set to 3.0 for the first frame and 1.0 for the last frame. The resolution and frame rate are set to 576×1024 and 7 fps, respectively, following the default settings. The decoding chunk size is set to 8 to balance generation quality and computational cost.

3.2. Hallucinating Flows

After generating target images from each source image using the image-to-video generation model, we obtain sets of source-target pairs for triplet construction. The temporary paired data P can be represented as:

$$P_s = \{(I_s, I_t) : t \in \{1, 2, \ldots, T\}\}, \tag{2}$$

where I_s represents the source image and I_t denotes the generated target image at frame t. From these temporary pairs, we construct complete training triplets for network training, as illustrated in Figure 2.

Triplet creation. To obtain the final training triplets for the two-stream video SOD network, we estimate optical flows from the temporary paired data P and leverage the corresponding ground-truth masks. For optical flow estimation, we use a pre-trained RAFT model [56] while maintaining the original image resolution. The final training triplets Q can be obtained as:

$$Q_s = \{(I_s, F_{s\to t}, M_s) : t \in \{1, 2, \ldots, T\}\}, \tag{3}$$

where $F_{s\to t}$ indicates the estimated optical flow map from the source frame s to target frame t, and M_s represents the ground-truth saliency mask corresponding to the source image I_s. Since optical flows are computed using the source image as the reference frame, the RGB images, segmentation masks, and optical flows maintain perfect spatial alignment. The triplet data Q is constructed for each static image

Table 1. Training data comparison highlighting the scale advantages of diffusion-based data generation.

Dataset	#Triplets	#Visual Contexts
DAVIS 2016 [45]	2,079	30
DAVSOD [18]	7,183	61
DUTS-Video	218,008	15,572

separately, and therefore, from N source images, we can obtain NT complete training triplets.

Dataset construction. We apply our triplet creation protocol to the DUTS [57] dataset, which consists of 10,553 training images and 5,019 testing images. Since these images are used solely for generating training data, we utilize the entire dataset (15,572 images total) for our pipeline. For each static image, we generate $T = 14$ target frames using video diffusion models, then extract optical flows between the source and each target frame, resulting in 218,008 training triplets (15,572×14). As shown in Table 1, this represents a substantial increase in scale compared to existing video SOD datasets: our DUTS-Video dataset contains 30× more triplets than DAVSOD [18] (7,183 triplets) and 105× more than DAVIS 2016 [45] (2,079 triplets), while providing 255× more visual contexts (15,572) than DAVSOD (61) and 519× more than DAVIS (30). These generated triplets provide complete image-flow-mask supervision and are combined with conventional video training data to enhance data diversity and provide realistic motion supervision for flow-guided video SOD training.

3.3. Segmentation Model

We train a video SOD network using our generated image-flow-mask triplets in a fully supervised manner. Each training triplet consists of an RGB image, its corresponding optical flow map, and ground-truth saliency mask, enabling end-to-end supervised learning. Following conventional flow-guided approaches, we employ a two-stream architecture that processes appearance and motion information separately, as illustrated in Figure 3.

Architecture details. Our network employs dual MiT [62] encoders: one for RGB appearance features and another for optical flow motion features. Multi-level features from both streams are fused through CBAM attention blocks [61], then progressively decoded using feature interpolation with skip connections. The decoder outputs binary segmentation masks for salient object detection, supervised by ground-truth annotations.

Training configuration. The network is trained in a fully supervised fashion using our triplet-based training data. We employ cross-entropy loss and Adam optimizer [32] with learning rate 1e-5, batch size 16, and input resolution 512×512. Training is performed on two GeForce RTX

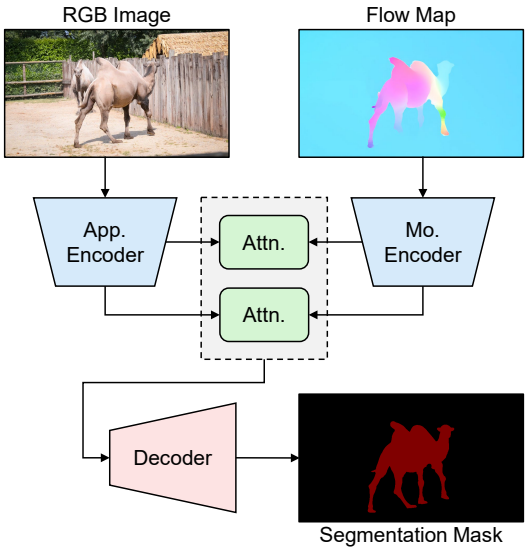


Figure 3. Two-stream network architecture for video SOD. The model processes RGB images and optical flow maps through separate encoders, fuses multi-level features via attention blocks, and outputs salient object masks.

TITAN GPUs for approximately two days. We combine real video triplets with our generated training triplets: real data includes DAVIS 2016 [45] and DAVSOD [18] training sets, while synthetic data comes from our pipeline applied to DUTS images. Training triplets are randomly shuffled with a 2:1:1 mixture ratio for generated data, DAVIS 2016, and DAVSOD, respectively, ensuring balanced exposure to both diverse synthetic motion patterns and real video dynamics while maintaining complete supervised learning.

4. Experiments

We evaluate TransFlow on multiple video SOD benchmarks and conduct ablation studies to validate our approach. All inference experiments are performed on a single GeForce RTX 2080 Ti GPU.

4.1. Datasets

We utilize existing image and video SOD datasets for training and evaluation. For video datasets, optical flows are precomputed between consecutive frames using RAFT [56] and stored alongside RGB images and segmentation masks.

DAVIS 2016 [45]. This primary video SOD benchmark contains 50 videos with frame-level annotations, split into 30 training and 20 validation videos. We use the training set for model training and the validation set for evaluation.

FBMS [40]. This dataset contains 59 videos (29 training, 30 testing) with sparse frame annotations. Following standard video SOD protocols, we use only the test set for evaluation.

DAVSOD [18]. A recent video SOD benchmark compris-

Table 2. Quantitative evaluation on the DAVIS 2016 validation set, FBMS test set, DAVSOD test set, and ViSal dataset.

Method	DAVIS 2016			FBMS			DAVSOD			ViSal			Average		
	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$
SSAV [18]	89.3	86.1	2.8	87.9	86.5	4.0	72.4	60.3	9.2	94.3	93.9	2.0	86.0	81.7	4.5
F^3Net [60]	85.0	81.9	4.1	85.3	81.9	6.8	68.9	56.4	11.7	87.4	90.7	4.5	81.7	77.7	6.8
MINet [43]	86.1	83.5	3.9	84.9	81.7	6.7	70.4	58.2	10.3	90.3	91.1	4.1	82.9	78.6	6.3
GateNet [66]	86.9	84.6	3.6	85.7	83.2	6.5	70.1	57.8	10.4	92.1	92.8	3.9	83.7	79.6	6.1
PCSA [21]	90.2	88.0	2.2	86.6	83.1	4.1	74.1	65.5	8.6	94.6	94.0	1.7	86.4	82.7	4.2
DFNet [67]	-	89.9	1.8	-	83.3	5.4	-	-	-	-	92.7	1.7	-	-	-
3DC-Seg [39]	-	91.8	1.5	-	84.5	4.8	-	-	-	-	92.2	1.9	-	-	-
CASNet [30]	87.3	86.0	3.2	85.6	86.3	5.6	69.4	-	8.9	82.0	84.7	2.9	81.1	-	5.2
FSNet [29]	92.0	90.7	2.0	89.0	88.8	4.1	77.3	68.5	7.2	-	-	-	-	-	-
CFAM [6]	91.8	90.9	1.5	90.9	91.5	2.6	75.3	66.2	8.3	94.7	95.1	1.3	88.2	85.9	3.4
UFO [55]	91.8	90.6	1.5	89.1	88.8	3.1	-	-	-	95.9	95.1	1.3	-	-	-
DBSNet [19]	92.4	91.4	1.4	88.2	88.5	3.8	77.8	68.8	7.6	93.1	92.8	2.0	87.9	85.4	3.7
HFAN [44]	93.4	92.9	**0.9**	87.5	84.9	3.3	75.3	68.0	7.0	94.1	93.5	1.1	87.6	84.8	3.1
TMO [11]	92.8	92.0	**0.9**	88.6	88.2	3.1	76.7	70.8	7.2	94.2	94.7	**1.0**	88.1	86.4	3.1
OAST [54]	93.5	92.6	1.1	91.7	**91.9**	**2.5**	78.6	71.2	7.0	94.8	95.0	**1.0**	89.7	87.7	2.9
TGFormer [20]	93.2	92.2	1.1	91.6	**91.9**	2.6	79.8	72.8	**6.5**	95.2	95.5	1.1	90.0	88.1	**2.8**
SimulFlow [28]	93.7	93.6	**0.9**	-	-	-	77.1	72.2	6.9	94.6	94.3	1.2	-	-	-
TransFlow	**94.5**	**93.9**	1.0	**92.6**	90.6	2.8	**80.3**	**73.2**	6.6	**96.2**	**96.6**	**1.0**	**90.9**	**88.6**	2.9

ing 61 training and 35 test videos with complete frame-level annotations. We use the training set for model training and the test set for evaluation.

ViSal [58]. This dataset contains 17 videos with 193 annotated frames. Since it lacks a defined train/test split, we use it solely for evaluation.

DUTS [57]. A large-scale image SOD dataset containing 10,553 training and 5,019 test images with pixel-level annotations. We use the entire dataset (15,572 images) for generating our DUTS-Video training data.

DUTS-Video. Our generated dataset created by applying our motion generation pipeline to DUTS images. From 15,572 static images, we generate 218,008 image-flow-mask triplets with semantically-aware motion patterns. This dataset is used exclusively for training TransFlow, enabling the model to benefit from realistic motion cues derived from abundant image data, effectively addressing the video data scarcity challenge.

4.2. Evaluation Metrics

We adopt three standard evaluation metrics for quantitative validation: $\mathcal{S}$-measure [17], $\mathcal{F}$-measure, and $\mathcal{M}$-measure.

$\mathcal{S}$-measure. The $\mathcal{S}$-measure evaluates structural similarity between predicted saliency maps and ground truth masks:

$$\mathcal{S} = \alpha\mathcal{S}_o + (1-\alpha)\mathcal{S}_r \,, \tag{4}$$

where $\mathcal{S}_o$ and $\mathcal{S}_r$ represent object-aware and region-aware structural similarity, respectively. We set $\alpha = 0.5$ following standard practice.

$\mathcal{F}$-measure. The $\mathcal{F}$-measure computes the weighted harmonic mean of precision and recall:

$$\mathcal{F} = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \,, \tag{5}$$

where $\beta^2 = 0.3$ following [1].

$\mathcal{M}$-measure. The $\mathcal{M}$-measure computes the mean absolute error between predictions and ground truth:

$$\mathcal{M} = \frac{1}{H \times W} \sum_{p=1}^{H} \sum_{q=1}^{W} |M_{pred}^{p,q} - M_{gt}^{p,q}| \,, \tag{6}$$

where M_{pred} and M_{gt} denote predicted and ground truth masks, respectively.

4.3. Quantitative Results

We evaluate TransFlow against state-of-the-art methods on four standard benchmarks: DAVIS 2016 [45] validation set, FBMS [40] test set, DAVSOD [18] test set, and ViSal [58] dataset. As shown in Table 2, TransFlow achieves the highest $\mathcal{S}$-measure [17] scores across all benchmarks, demonstrating superior structural similarity preservation. For $\mathcal{F}$-measure, our method ranks first on three datasets and remains competitive on the fourth, while maintaining comparable $\mathcal{M}$-measure performance.

The consistent improvements across diverse datasets validate that our semantic motion generation approach meaningfully enhances video SOD performance. These results demonstrate that addressing training data scarcity through

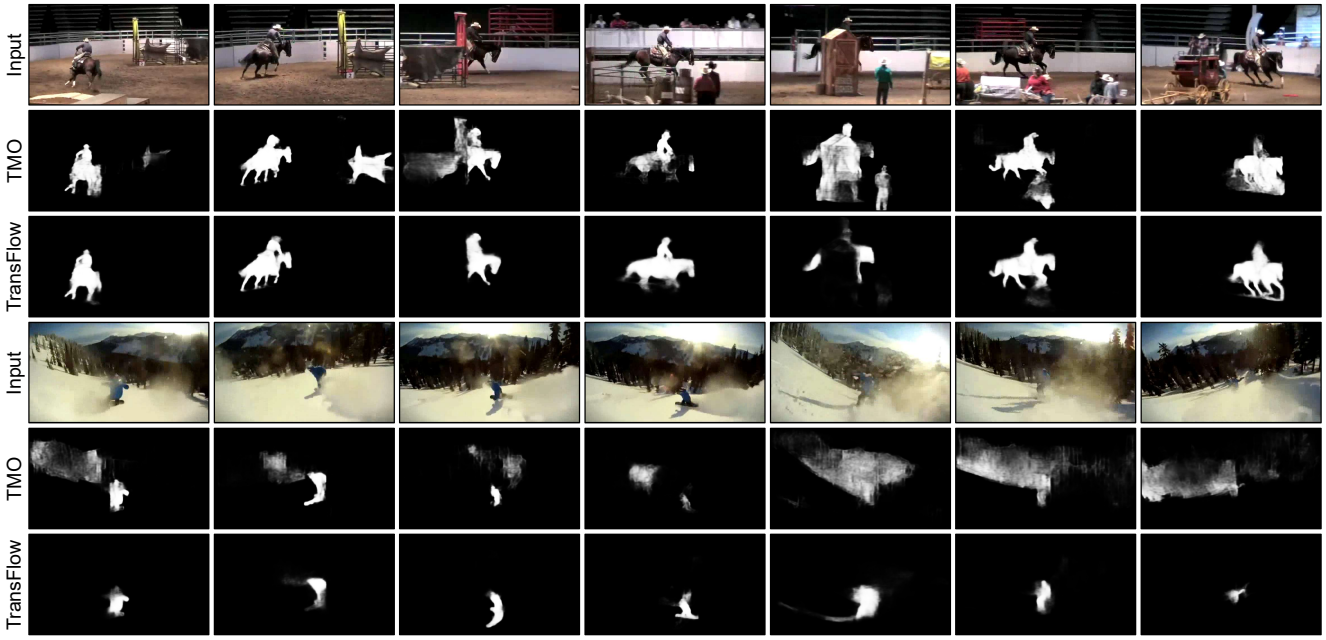


Figure 4. Qualitative comparison under challenging scenarios.

Table 3. Ablation study on the training protocol.

Training	DAVIS 2016			FBMS			DAVSOD			ViSal			Average		
	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$
Real	93.1	91.6	1.3	85.8	85.5	5.3	76.0	68.7	7.8	94.5	94.4	1.8	87.4	85.1	4.1
Synthetic	91.7	89.6	2.0	91.0	89.0	2.9	78.1	70.1	7.8	96.1	96.4	1.1	89.2	86.3	3.5
Mixed	94.5	93.9	1.0	92.6	90.6	2.8	80.3	73.2	6.6	96.2	96.6	1.0	90.9	88.6	2.6

large-scale synthesis is more effective than architectural innovations for advancing video SOD capabilities. Our findings suggest that data abundance, rather than model sophistication, represents the primary bottleneck in current video SOD development.

4.4. Qualitative Results

Figure 4 compares TransFlow with TMO [11] on challenging scenarios featuring complex background clutter and severe occlusion. While TMO produces fragmented segmentations and misses salient regions, TransFlow delivers complete and accurate object boundaries despite significant visual ambiguities and distractors.

This performance advantage stems from our large-scale synthetic training data, which exposes the model to diverse appearance-motion combinations unavailable in limited real datasets. The significant increase in training triplets compared to existing video SOD datasets enables robust feature learning and superior generalization to challenging real-world scenarios, demonstrating that synthetic data generation can effectively overcome the fundamental data scarcity bottleneck in video SOD.

4.5. Analysis

We conduct experiments to evaluate our approach, focusing on training protocols, data characteristics, architectural choices, and limitations to understand the mechanisms behind performance improvements.

Training protocol. Table 3 evaluates different training data configurations. Training exclusively on our DUTS-Video dataset outperforms training on real video data (DAVIS 2016 [45] and DAVSOD [18]) across all benchmarks except the DAVIS 2016 validation set. This demonstrates that semantically-aware motion patterns generated through our knowledge transfer approach provide more effective training supervision than limited real video data. By leveraging motion knowledge from pre-trained video diffusion models, our method generates realistic training data that enhances video SOD model development. Combining real video data with DUTS-Video achieves optimal performance across all datasets and metrics, confirming that abundant training data with realistic motion cues significantly enhances video SOD model development.

Comparison to real data. Our approach generates diverse image-flow-mask triplets by creating video sequences from

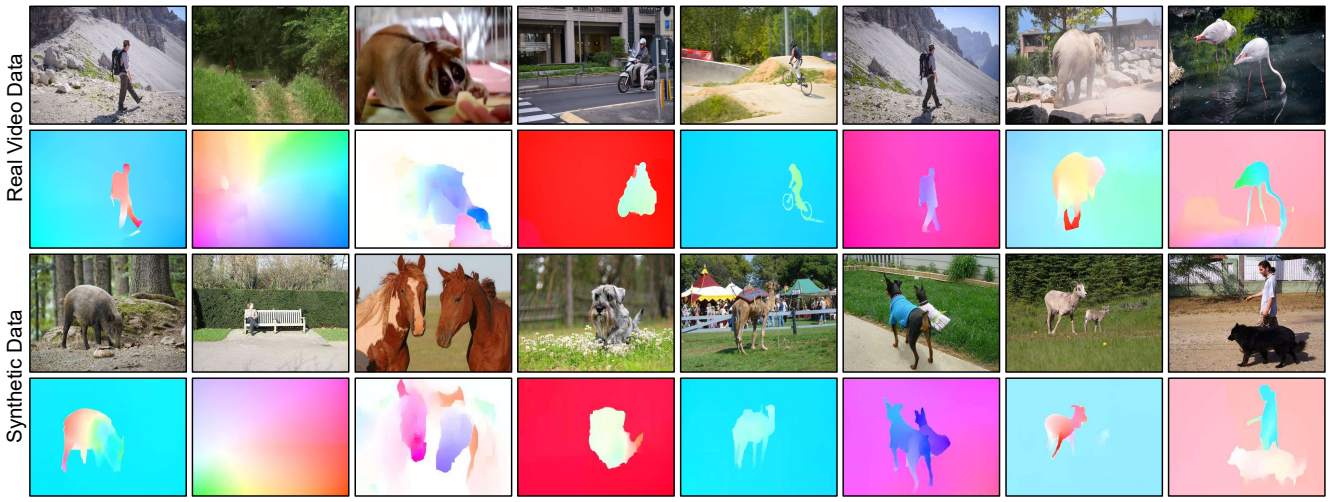


Figure 5. Aligned qualitative comparison between real video data and our synthetic data.

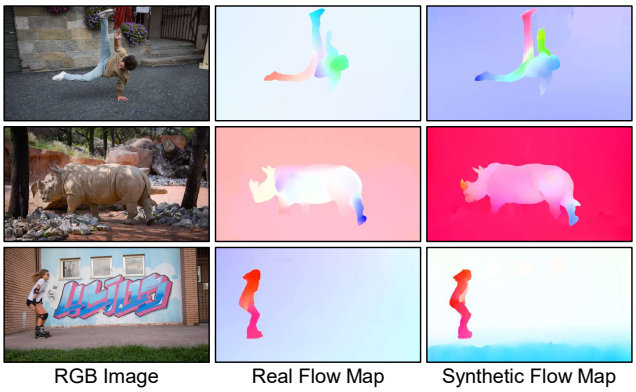


Figure 6. Direct qualitative comparison of actual and synthetic flow maps, both derived from the same source image.

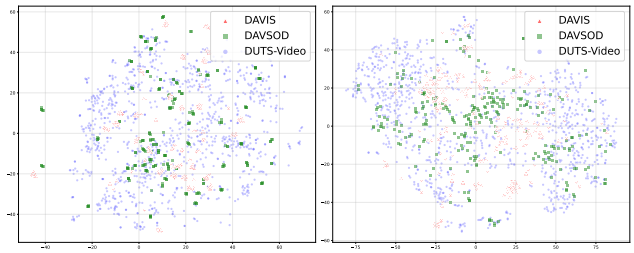


Figure 7. t-SNE comparison of feature distributions from real and synthetic datasets (left: RGB images, right: optical flow maps).

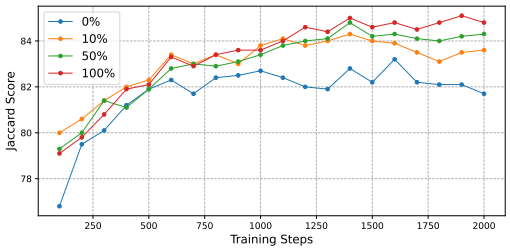


Figure 8. Jaccard score across training steps for different ratios of synthetic data, evaluated on the DAVIS 2016 validation set.

static images and extracting motion patterns from the generated frames. Figure 5 shows that our synthetic data encompass diverse optical flow patterns, from clear object-shaped flows to ambiguous motion representations, providing comprehensive coverage for robust training. Figure 6 further validates our approach through direct comparison with real flows from identical source images, demonstrating that our generated flows capture plausible object movements suitable for video SOD training. This highlights the effectiveness of transferring motion knowledge from video diffusion models to generate semantically-aware optical flows.

Data distribution. Figure 7 visualizes feature embeddings extracted using ResNet-18 [24] and projected via t-SNE to compare data distributions across datasets. Real video datasets exhibit distinct, limited distributions, explaining the poor cross-dataset generalization observed in Table 3. In contrast, our synthetic data spans a significantly broader feature space, providing comprehensive appearance and motion diversity that enables robust generalization across various domains and challenging scenarios. This broader feature space reflects the rich semantic motion patterns embedded in our generated training data.

Training statistics. We systematically analyze the effect of synthetic training data volume on both convergence speed and segmentation performance. As shown in Figure 8, models trained with a higher proportion of synthetic data require slightly more training steps to converge, yet consistently achieve superior Jaccard scores compared to those trained with less. These results highlight that, while increased synthetic data may marginally slow convergence, it provides a clear and substantial benefit to overall performance.

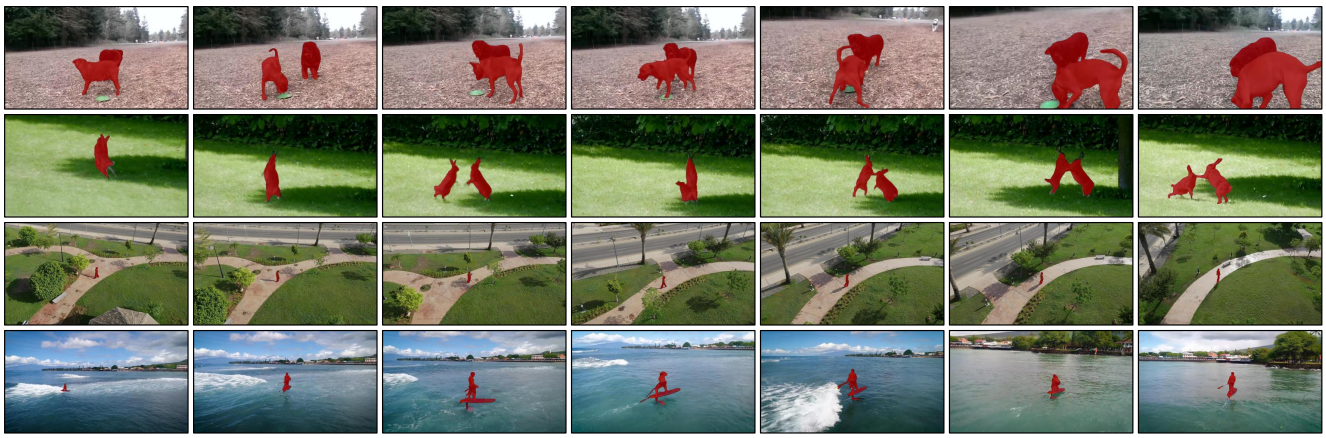


Figure 9. Qualitative results on the general video cases.

Table 4. Ablation study on the backbone network.

Backbone	fps	#Param	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$
MiT-b0	61.1	13.8M	89.4	87.0	3.5
MiT-b1	45.5	33.6M	89.8	87.3	3.2
MiT-b2	29.5	55.7M	90.9	88.6	2.6

Table 5. Ablation study on the input resolution.

Resolution	fps	$\mathcal{S}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{M}\downarrow$
256×256	55.0	90.0	87.5	3.0
384×384	44.2	90.2	87.8	2.9
512×512	29.5	90.9	88.6	2.6

Backbone network. Table 4 compares different backbone architectures across four benchmarks. Our default MiT-b2 [62] backbone achieves optimal performance while maintaining efficient inference speed and parameter count, making it suitable for practical applications. The consistent performance scaling with backbone capacity demonstrates the stability and effectiveness of our training methodology across different architectural choices.

Input resolution. Table 5 examines performance-efficiency trade-offs across different input resolutions. Lower resolutions provide faster inference with minimal accuracy degradation, while higher resolutions maximize precision. This flexibility allows adaptation to various computational constraints and application requirements without compromising the core method's effectiveness.

General video cases. Figure 9 shows the results on general videos from the MOSE [15] validation set and the LVOS [27] test set. We sample scenarios where video saliency is valid. The upper two are from the MOSE validation set, and the lower two are from the LVOS test set. We do not perform additional training on these datasets and directly test the trained model on them. The results

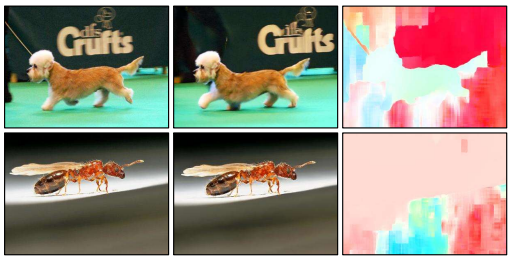


Figure 10. Visualization of the limitations of our simulated flows.

demonstrate that TransFlow is effective for handling general video cases, further validating the motion knowledge transfer from video diffusion models.

Limitations. Our synthetic data generation approach has several limitations. The generated videos occasionally exhibit discontinuous motion dynamics, leading to checkerboard artifacts in the optical flow, as shown in Figure 10. These artifacts likely arise from patch-level content copying inherent to the diffusion model. Furthermore, since only RGB frames are synthesized, corresponding transformed masks cannot be generated, which limits the diversity of source-target combinations. Addressing these issues presents a promising direction for future work, such as enabling the joint generation of images and masks.

5. Conclusion

In this study, we investigate the use of large-scale image data to train flow-guided models. Rather than relying on source image distortion methods that fail to accurately mimic the optical flows of real video data, we demonstrate that novel image generation through image-to-video generation provides an effective solution. For the video SOD task, our approach sets a new state-of-the-art performance across all benchmark datasets, surpassing existing methods.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. v

[2] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 158–177. Springer, 2020. ii

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. ii, iii

[4] Fred L Bookstein and WDK Green. A thin-plate spline and the decomposition of deformations. *Mathematical Methods in Medical Imaging*, 2(14-28):3, 1993. ii

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024. ii

[6] Yi-Wen Chen, Xiaojie Jin, Xiaohui Shen, and Ming-Hsuan Yang. Video salient object detection via contrastive features and attention modules. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1320–1329, 2022. v

[7] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. ii

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. ii

[9] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014. ii

[10] Suhwan Cho, Heansung Lee, Minhyeok Lee, Chaewon Park, Sungjun Jang, Minjung Kim, and Sangyoun Lee. Tackling background distraction in video object segmentation. In *European Conference on Computer Vision*, pages 446–462. Springer, 2022. ii

[11] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Chaewon Park, Donghyeong Kim, and Sangyoun Lee. Treating motion as option to reduce motion dependency in unsupervised video object segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5140–5149, 2023. v, vi

[12] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Dogyoon Lee, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Dual prototype attention for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19238–19247, 2024. ii

[13] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. iii

[14] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703, 2023. i

[15] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20224–20234, 2023. i, viii

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. ii

[17] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. v

[18] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8554–8564, 2019. iv, v, vi

[19] Jiaqing Fan, Tiankang Su, Kaihua Zhang, and Qingshan Liu. Bidirectionally learning dense spatio-temporal feature propagation network for unsupervised video object segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3646–3655, 2022. v

[20] Jiaqing Fan, Tiankang Su, Kaihua Zhang, Bo Liu, and Qingshan Liu. Temporally efficient gabor transformer for unsupervised video object segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3394–3402, 2023. v

[21] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Ming-Ming Cheng, and Shao-Ping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10869–10876, 2020. v

[22] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025. ii

[23] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. ii

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. vii

[25] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. iii

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. ii

[27] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13480–13492, 2023. viii

[28] Lingyi Hong, Wei Zhang, Shuyong Gao, Hong Lu, and WenQiang Zhang. Simulflow: Simultaneously extracting feature and identifying target for unsupervised video object segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7481–7490, 2023. v

[29] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4922–4933, 2021. ii, v

[30] Yuzhu Ji, Haijun Zhang, Zequn Jie, Lin Ma, and QM Jonathan Wu. Casnet: A cross-attention siamese network for video salient object detection. *IEEE transactions on neural networks and learning systems*, 32(6):2676–2690, 2020. v

[31] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. iii

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. iv

[33] Pika Labs. Pika 1.0, 2023. https://pika.art/. ii

[34] Minhyeok Lee, Suhwan Cho, Seunghoon Lee, Chaewon Park, and Sangyoun Lee. Unsupervised video object segmentation via prototype memory network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5924–5934, 2023. ii

[35] Minhyeok Lee, Suhwan Cho, Dogyoon Lee, Chaewon Park, Jungho Lee, and Sangyoun Lee. Guided slot attention for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3807–3816, 2024. ii

[36] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7274–7283, 2019. ii

[37] Xiangtai Li, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2024. i

[38] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pages 56–74. Springer, 2025. ii

[39] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. *arXiv preprint arXiv:2008.11516*, 2020. ii, v

[40] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6): 1187–1200, 2013. iv, v

[41] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7376–7385, 2018. ii

[42] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9226–9235, 2019. ii

[43] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9413–9422, 2020. v

[44] Gensheng Pei, Fumin Shen, Yazhou Yao, Guo-Sen Xie, Zhenmin Tang, and Jinhui Tang. Hierarchical feature alignment network for unsupervised video object segmentation. In *European Conference on Computer Vision*, pages 596–613. Springer, 2022. v

[45] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. iv, v, vi

[46] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2663–2672, 2017. ii

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. ii

[48] runway. Gen-2, 2023. https://runwayml.com/research/gen-2. ii

[49] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 629–645. Springer, 2020. ii

[50] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021. ii

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

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. ii

[53] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. iii

[54] Tiankang Su, Huihui Song, Dong Liu, Bo Liu, and Qingshan Liu. Unsupervised video object segmentation with online adversarial self-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 688–698, 2023. v

[55] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, 26:313–325, 2023. v

[56] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. iii, iv

[57] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. iv, v

[58] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015. v

[59] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017. ii

[60] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12321–12328, 2020. v

[61] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. iv

[62] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. iv, viii

[63] Shu Yang, Lu Zhang, Jinqing Qi, Huchuan Lu, Shuo Wang, and Xiaoxing Zhang. Learning motion-appearance co-attention for zero-shot video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1564–1573, 2021. ii

[64] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. ii

[65] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. ii

[66] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 35–51. Springer, 2020. v

[67] Mingmin Zhen, Shiwei Li, Lei Zhou, Jiaxiang Shang, Haoan Feng, Tian Fang, and Long Quan. Learning discriminative feature with crf for unsupervised video object segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 445–462. Springer, 2020. v

[68] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13066–13073, 2020. ii