Video Object Segmentation-aware Video Frame Interpolation

Jun-Sang Yoo
Korea University
junsang7777@korea.ac.kr

Hongjae Lee
Korea University
jimmy9704@korea.ac.kr

Seung-Won Jung*
Korea University
swjung83@korea.ac.kr

Abstract

Video frame interpolation (VFI) is a very active research topic due to its broad applicability to many applications, including video enhancement, video encoding, and slow-motion effects. VFI methods have been advanced by improving the overall image quality for challenging sequences containing occlusions, large motion, and dynamic texture. This mainstream research direction neglects that foreground and background regions have different importance in perceptual image quality. Moreover, accurate synthesis of moving objects can be of utmost importance in computer vision applications. In this paper, we propose a video object segmentation (VOS)-aware training framework called VOS-VFI that allows VFI models to interpolate frames with more precise object boundaries. Specifically, we exploit VOS as an auxiliary task to help train VFI models by providing additional loss functions, including segmentation loss and bi-directional consistency loss. From extensive experiments, we demonstrate that VOS-VFI can boost the performance of existing VFI models by rendering clear object boundaries. Moreover, VOS-VFI displays its effectiveness on multiple benchmarks for different applications, including video object segmentation, object pose estimation, and visual tracking. The code is available at https://github.com/junsang7777/VOS-VFI

1. Introduction

Video frame interpolation (VFI) is a low-level video processing task synthesizing intermediate frames between consecutive frames to increase the frame rate. VFI has been widely used in many applications, including slow motion generation [44, 46, 2, 27], video restoration [67, 31], video compression [65], and novel view synthesis [17, 29]. Although VFI has been extensively studied in past decades, there is still room for improvement because moving objects, occlusions, and cluttered backgrounds make it challenging.

Like other vision tasks, deep learning-based methods dominate VFI, which can be classified into kernel-based [35, 46, 15, 55] and flow-based [14, 27, 70, 32, 40] approaches. The former learns interpolation kernels for two consecutive frames as an output of a convolutional neural network (CNN) to synthesize an intermediate frame; whereas the latter finds optical flow between two frames in the pixel-space or feature-space and generates an intermediate frame by motion compensated prediction. Both approaches have been advanced by adopting novel architectures or algorithms, including coarse-to-fine architectures [56, 72, 32], attention mechanisms [11, 30], de-

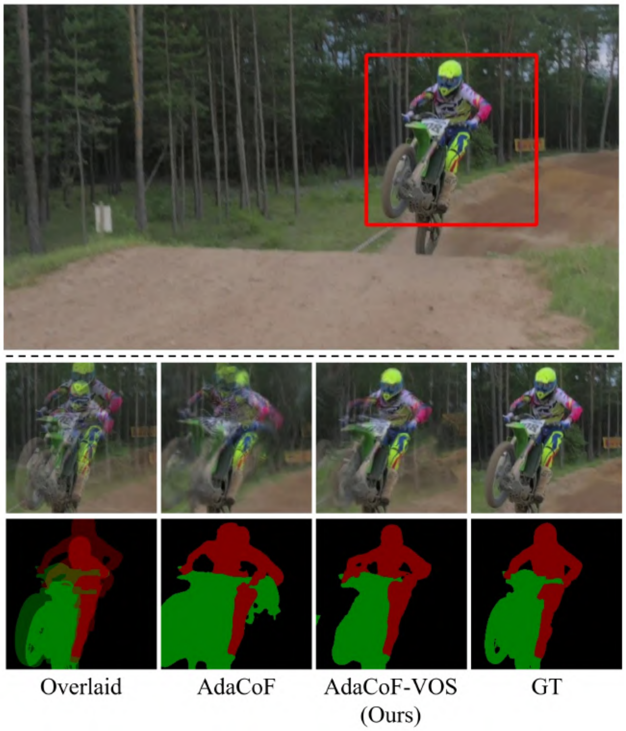

Figure 1: Examples of video frame interpolation and video object segmentation results on the AdaCoF baseline [35]. The proposed VOS-VFI is a training framework that can improve the performance of the baseline model without increasing the number of parameters and inference time.

[*] Corresponding author.

formable convolutions [21, 35], and Transformers [55, 40].

Although we have witnessed significant performance improvements in several VFI benchmarks [11, 71, 57, 50], VFI methods have been competing to improve their performance in terms of the overall quality of interpolated images, *e.g.*, PSNR and SSIM [24]. This bias in the benchmarks neglects the fact that foreground and background regions have different levels of importance in perceptual image quality [53]. In addition, VFI can contribute to the performance improvement of vision applications, such as video object tracking and object pose estimation. However, it is undiscovered whether VFI methods depicting high performance in global image quality perform consistently well in vision applications.

In this paper, we propose to perform video object segmentation (VOS) as an auxiliary task during the training of the VFI model such that the interpolated frames have clear object boundaries without artifacts. Specifically, the proposed VOS-aware VFI, called VOS-VFI, performs VOS using the existing and interpolated frames and uses the accuracy of VOS as an auxiliary loss term. Moreover, the VOS accuracy is measured in both forward and backward directions, and the forward-backward consistency is further enforced considering the temporal coherence of VFI.

Extensive experimental results demonstrate that the proposed VOS-VFI can be applied to existing state-of-the-art VFI models, including AdaCoF [35], CDFI [15], and IFRNet [32], making them produce interpolated frames with sharper and clearer objects. These improved results lead to the performance improvement of not only VOS but also other related vision tasks.

The main contributions are summarized as follows:

- We propose a new training framework for VFI called VOS-VFI. To the best of our knowledge, VOS-VFI is the first work that exploits VOS to assist in training VFI models.
- We design segmentation and bi-directional consistency loss terms of VOS that allow VFI models to render frames with precise object boundaries with temporal consistency.
- Comprehensive experimental results demonstrate the effectiveness of VOS-VFI on several vision applications, including video object segmentation, video object tracking, and object pose estimation.

2. Related Works

2.1. Video Frame Interpolation

Deep learning-based VFI methods can be categorized into kernel-based and flow-based approaches. The kernel-based approach uses CNNs to estimate interpolation kernels to be applied to the existing frames for the intermediate frame synthesis. Because the kernel-based approach does not explicitly perform motion estimation and compensation between frames, large-size and pixel-adaptive interpolation kernels are typically required, making kernel prediction challenging [45]. To this end, Niklaus *et al.* [46] proposed to estimate pairs of 1D separable kernels for each pixel such that large kernels can be estimated without requiring excessive memory. Furthermore, deformable convolutions [13] have been extensively used for VFI by simultaneously predicting the position and weights of the kernels [10, 35, 54]. A recent Transformer-based network called VFIT [55] shows state-of-the-art performance by fully exploiting long-range dependencies with self-attention operations.

The flow-based approach is inspired by traditional VFI algorithms [12, 26] that perform motion compensated prediction for image synthesis. Instead of performing motion estimation and compensation by handcrafted algorithms, deep learning-based methods employ flow estimation networks [44, 70] and differentiable warping layers [16, 64] for VFI. These methods typically synthesize two frames along the forward and backward directions and combine them using image synthesis networks [49, 44, 15, 14]. To deal with a wider range of motions, various VFI-tailored motion estimation methods have been proposed, including bidirectional flow estimation [27, 56, 14] and asymmetric bilateral motion estimation [49]. A recent network called VFIFormer [40] applies Transformer blocks to the warped frames and features to model long-range dependencies and demonstrates state-of-the-art performance. Several hybrid methods [2, 3] that first warp images and then apply local interpolation kernels have also been introduced to take advantage of kernel-based and flow-based approaches.

2.2. Video Object Segmentation

Existing VOS methods can be classified into semi-supervised and unsupervised approaches, where the former segments objects using a given annotation in the first frame, and the latter extracts objects using visual saliency or motion patterns without requiring annotations. Earlier deep learning-based VOS methods paid attention to online learning [9, 61] that fine-tunes the VOS network using the predicted object masks of the given video sequence. These methods typically require hyper-parameter settings for each sequence and high computational costs for the inference. To solve this problem, offline learning methods using pixel or feature-level matching [8, 47], graph optimization [41, 63], and optical flow warping [37, 68] have been proposed and have shown superior performance over online learning methods. The interested reader can refer to recent articles [18, 50] for a structured literature review of VOS.

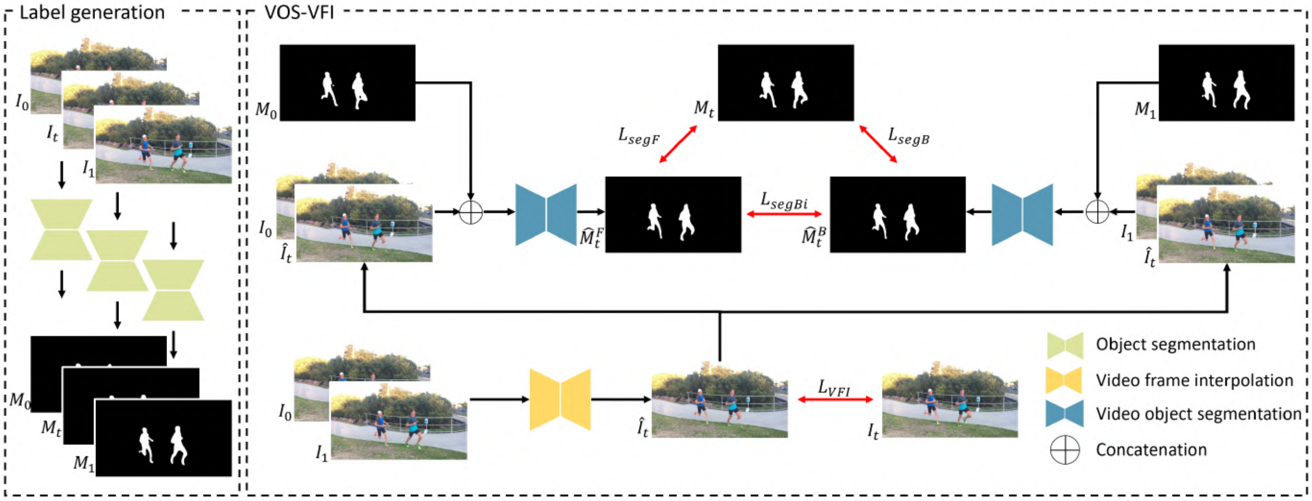

Figure 2: Overall framework of VOS-VFI. Given a triplet of images $\{I_0, I_t, I_1\}$, a triplet of object masks $\{M_0, M_t, M_1\}$ is first constructed. The result of the VFI model, *i.e.*, $\hat{I}_t$, is then fed to the VOS model that takes $\{I_0, \hat{I}_t, M_0\}$ or $\{I_1, \hat{I}_t, M_1\}$ as input. The results of the VOS model, *i.e.*, $\hat{M}_t^F$ and $\hat{M}_t^B$, are compared with each other and with M_t to guide the VFI model training.

In this paper, we do not propose a new VOS method but use VOS as an auxiliary task to help VFI. In line with our objective, several related studies that exploit the segmentation task to assist other vision tasks have been introduced in the literature. For example, Gidaris *et al.* [19] used two adaptive network modules to improve object detection performance by integrating regional object features with semantic segmentation-aware features. Harley *et al.* [22] extracted foreground-focused local features using segmentation-aware CNNs, resulting in significant performance improvements in optical flow estimation and semantic segmentation tasks. For the image restoration tasks, including deraining and denoising, segmentation loss terms defined between the features extracted using a pre-trained segmentation network [38] demonstrated effectiveness [62, 74].

3. Proposed Method

Given two frames I_0 and I_1 and a time $t \in (0, 1)$, VFI aims to synthesize the intermediate frame I_t. Most existing VFI models are trained on pixel-level loss, such as the L1 loss, to improve the overall quality of the interpolated images. In this paper, we claim that it is necessary to pay special attention to moving objects for VFI. To realize our objective of object-aware VFI, we present VOS-VFI that integrates a VOS model during the training of a VFI model. Fig. 2 illustrates the overall framework of VOS-VFI.

3.1. Segmentation Label Generation

The training of VOS-VFI requires object annotations for a sufficient amount of video frames. One can consider using existing VOS datasets, such as DAVIS [50, 51] and YouTubeVOS [69]; however, they are small in size and limited in content diversity. On the contrary, existing VFI datasets support a large number of video clips with diverse content for both indoor and outdoor scenes, *e.g.*, Vimeo90K [71] contains 73,171 frame triplets from 14,777 video clips. Therefore, we decide to obtain object annotations from the VFI dataset [71] by applying a pre-trained off-the-shelf segmentation model [7]. Specifically, for an image triplet $\{I_0, I_t, I_1\}$, where $t = 0.5$, we construct a triplet of object masks $\{M_0, M_t, M_1\}$. The object masks for the existing frames, *i.e.*, M_0 and M_1, are used as input for the VOS model, and the object mask of the ground-truth intermediate frame, *i.e.*, M_t, is used to supervise the training of the VFI model.

3.2. VOS-aware Training

The L1 loss has been standard for training the VFI model, which is prone to synthesizing blurry frames. Several alternative loss terms, including adversarial loss [20] and perceptual loss [28], have demonstrated effectiveness; however, it is still challenging to harmonize the existing and interpolated frames without flickering artifacts using these loss functions for the scenes with large object motions.

Humans focus more on the foreground than the background [53]. In addition, moving objects draw more attention than static objects [5]. Many video applications also require high-quality rendering of moving objects. Considering these conditions, we propose to use VOS as a provider for an auxiliary loss. Specifically, we adopt the VOS model $\mathcal{F}_{\Theta_{VOS}}$ that takes the previous frame and its predicted segmentation mask as well as the current frame as input and produces the segmentation map for the current frame as output. Most existing semi-supervised VOS models support this input-output configuration, and we used the off-the-shelf VOS model [8] in our experiment. Let the VFI model $\mathcal{F}_{\Theta_{VFI}}$ take two frames, *i.e.*, I_0 and I_1, as input and produce an interpolated frame $\hat{I}_t$ as output. Then, we feed the interpolated frame to the VOS model and obtain the segmentation map. By penalizing the difference between the predicted and ground-truth segmentation maps, we can make the VFI model synthesize images with accurate object boundaries.

The proposed training framework is not dependent on specific VFI model architectures. In Sec. 4, we will show the effectiveness of VOS-VFI on the three baselines, including AdaCoF [35], CDFI [15], and IFRNet [32]. In addition, because the VOS model is detached during evaluation, the test-time complexity of the adopted baselines is unchanged by VOS-VFI.

3.3. Objective Function

Let $L_{VOS-VFI}$ denote the total loss function, defined as follows:

$$L_{VOS-VFI} = L_{VFI} + \lambda_1 \left(L_{segF} + L_{segB}\right) + \lambda_2 L_{segBi}, \tag{1}$$

where λ_1 and λ_2 are weighting factors, and L_{VFI} represents the loss of the baseline model, which is different for the adopted baseline models [35, 15, 32]. VOS-VFI further applies three novel loss functions, L_{segF}, L_{segB}, and L_{segBi}, which will be detailed in the following subsections.

3.3.1 Segmentation Loss

To guide the VFI model to synthesize a foreground-focused interpolated frame, we perform VOS on the interpolated frames. The VOS model predicts the segmentation map of the interpolated frame using the previous frame and its segmentation map. In our training configurations with given image and mask triplets, *i.e.*, $\{I_0, I_t, I_1\}$ and $\{M_0, M_t, M_1\}$, we can apply the VOS model along two directions. Specifically, the VOS is performed in the forward direction, resulting in $\hat{M}_t^F = \mathcal{F}_{\Theta_{VOS}}\left(I_0, \hat{I}_t, M_0\right)$, and the backward direction, resulting in $\hat{M}_t^B = \mathcal{F}_{\Theta_{VOS}}\left(I_1, \hat{I}_t, M_1\right)$. Then, the VFI model is trained to minimize the difference between the target mask M_t and predicted masks $\hat{M}_t^F$ and $\hat{M}_t^B$. To this end, the segmentation loss functions along the forward and backward directions, *i.e.*, L_{segF} and L_{segB}, are defined as follows:

$$L_{segF} = CE\left(\hat{M}_t^F, M_t\right), \tag{2}$$

$$L_{segB} = CE\left(\hat{M}_t^B, M_t\right), \tag{3}$$

where CE measures the cross-entropy loss [73]. By penalizing the deviation of the prediction segmentation masks from the (pseudo) ground-truth segmentation mask, we can enforce the VFI model to synthesize frames with clear object boundaries.

3.3.2 Bi-directional Consistency Loss

Temporal consistency is of utmost importance in VFI because interpolated frames are inserted between existing frames. Since we have two predictions, *i.e.*, $\hat{M}_t^F$ and $\hat{M}_t^B$, for the target mask M_t, we can also penalize the difference between $\hat{M}_t^F$ and $\hat{M}_t^B$ to enforce the temporal coherence of VOS, eventually leading to the temporal coherence of VFI. To this end, the bi-directional segmentation consistency loss L_{segBi} is defined as follows:

$$L_{segBi} = CE\left(\hat{M}_t^F, \hat{M}_t^B\right). \tag{4}$$

Although optical flow provides dense correspondences, it is vulnerable to noise, occlusion, and brightness changes across frames. We instead enforce foreground-focused temporal consistency in the segmentation domain. As a result, it allows the VFI model to interpolate salient moving objects in the intermediate frames more precisely.

4. Experiments

We first provide the implementation details of VOS-VFI. We then evaluate the VFI performance of VOS-VFI on the three representative VFI models, *i.e.*, AdaCoF [35], CDFI [15], and IFRNet [32]. For notational simplicity, the baseline models trained with the proposed VOF-VFI are denoted with the suffix "-VOS," *e.g.*, CDFI-VOS. Next, we further show the effectiveness of VOS-VFI on other related tasks, *i.e.*, video object tracking and object pose estimation. Last, we conduct ablation studies to verify the effectiveness of different loss functions in VOS-VFI. More results and code can be found in the supplementary material.

4.1. Implementation Details

The training settings of the VFI models, including learning rates, batch size, and patch size, were derived from the public codes of the baseline models [35, 15, 32], and the weighting parameters in (1) were empirically chosen as $\lambda_1 = \lambda_2 = 0.1$. We used a pre-trained STCN model [8] as the VOS model and fixed it during the training of the VFI

Model	DAVIS 2016					DAVIS 2017				
	PSNR$_\uparrow$	†PSNR$_\uparrow$	$J\&F_\uparrow$	$J_\uparrow$	$F_\uparrow$	PSNR$_\uparrow$	†PSNR$_\uparrow$	$J\&F_\uparrow$	$J_\uparrow$	$F_\uparrow$
AdaCoF [35]	**25.11**	25.62	85.9	84.9	86.8	**26.23**	26.13	80.9	77.9	83.9
AdaCoF-VOS	25.03	**25.72**	**87.0**	**85.9**	**88.2**	26.21	**26.22**	**81.8**	**78.6**	**84.9**
CDFI [15]	25.68	25.74	86.9	85.6	88.3	26.71	26.24	81.4	78.3	84.6
CDFI-VOS	**25.75**	**25.80**	**87.7**	**86.4**	**88.9**	**26.79**	**26.30**	**82.0**	**78.9**	**85.3**
IFRNet [32]	26.70	25.91	87.3	86.4	88.3	27.57	26.44	81.1	77.9	84.3
IFRNet-VOS	**26.74**	**25.98**	**88.1**	**87.2**	**88.9**	**27.60**	**26.50**	**81.9**	**78.6**	**85.2**

Table 1: Quantitative results of the three baseline models trained with and without VOS-VFI. The performance is evaluated in terms of the PSNR and segmentation accuracy ($J\&F$, J, and F) on the DAVIS 2016 and 2017 datasets. † represents PSNR scores on the foreground objects obtained by masking out the background using the ground-truth segmentation maps.

Model	Vimeo90K (val)			UCF101 (val)			Xiph 2K			Xiph 4K		
	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$
AdaCoF [35]	5.180	4.104	4.765	7.272	5.695	4.018	4.043	4.637	4.470	5.217	5.817	4.349
AdaCoF-VOS	**5.153**	**4.094**	**4.771**	**7.237**	**5.678**	**4.026**	**4.014**	**4.616**	**4.476**	**5.203**	**5.809**	**4.353**
CDFI [15]	4.933	3.832	4.873	6.878	5.421	3.987	3.804	4.533	4.473	4.773	5.106	4.356
CDFI-VOS	**4.910**	**3.822**	**4.879**	**6.875**	**5.408**	**3.991**	**3.751**	**4.520**	**4.477**	**4.694**	**5.089**	**4.374**
IFRNet [32]	5.062	3.969	4.820	7.191	5.665	**4.023**	4.008	4.528	4.461	5.470	5.971	4.315
IFRNet-VOS	**5.021**	**3.935**	**4.824**	**7.115**	**5.617**	4.020	**3.974**	**4.503**	**4.480**	**5.448**	**5.959**	**4.331**

Table 2: Quantitative results of the three baseline models trained with and without VOS-VFI. The performance is evaluated in terms of the three representative image quality metrics on the Vimeo90K, UCF101, and Xiph datasets.

models. Since VOS is performed during VOS-VFI training, the complexity of the training inevitably increases. For example, when STCN [8] was used within the VOS-VFI training framework, the total training time of AdaCoF increased from 40h to 51h on an NVIDIA 3090 GPU, and the maximum memory consumption increased from 9.6GB to 10.1GB after applying the proposed training framework.

We obtained binary masks for the segmentation label generation, as described in Sec. 3.1, by applying a DeepLabV3 model [7] pre-trained on the COCO dataset [39] to the training images from Vimeo90K [71] and extracting the pixels belonging to foreground object classes. Consequently, the training dataset for VOS-VFI includes 51,312 frame triplets from Vimeo90K and their corresponding mask triplets.

4.2. Performance Evaluation on Multiple Tasks

4.2.1 VOS-VFI

We first evaluated the VFI performance on the DAVIS benchmarks [50, 51] that support performance evaluation in terms of both image quality and segmentation accuracy. DAVIS, one of the most widely-used VOS benchmarks, consists of two sets: (1) DAVIS 2016, which is an object-level annotated dataset (single object); and (2) DAVIS 2017, which is an instance-level annotated dataset (multiple objects). In addition to image-level PSNRs, PSNRs on foreground objects were measured by masking out the background using the ground-truth segmentation maps. The segmentation accuracy metrics, *i.e.*, region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$ [52], were also measured for performance comparisons.

Model	HD 544p		
	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$
AdaCoF [35]	5.611	5.037	4.708
AdaCoF-VOS	**5.581**	**4.992**	**4.740**
CDFI [15]	5.903	4.680	5.227
CDFI-VOS	**5.868**	**4.665**	**5.237**
IFRNet [32]	5.792	5.117	4.552
IFRNet-VOS	**5.648**	**5.086**	4.552

Table 3: Quantitative evaluation for 4× interpolation on the HD [3] benchmark.

As shown in Table 1, the proposed VOS-VFI improved

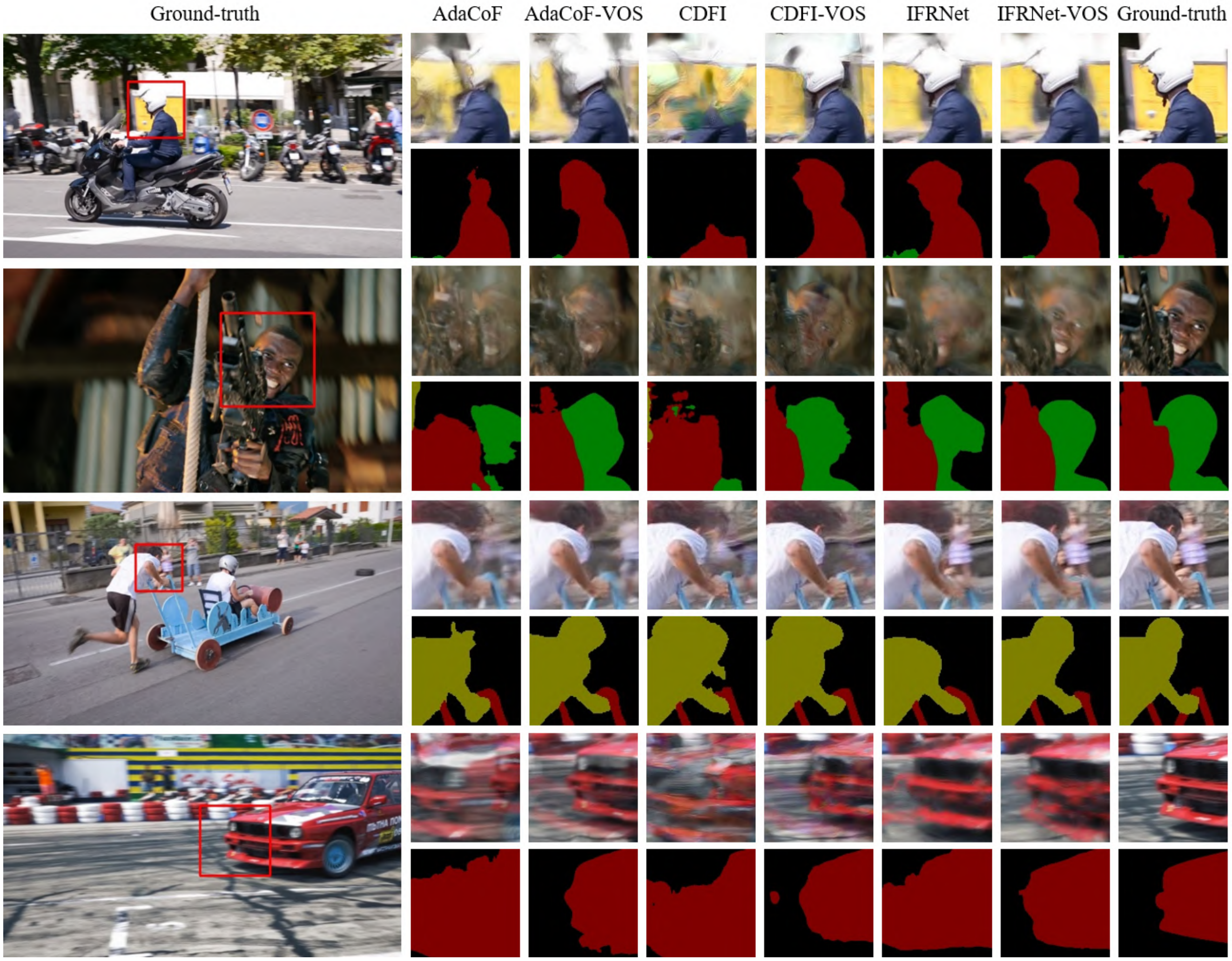

Figure 3: Qualitative performance comparisons of VOS-VFI. The images are from the validation dataset of DAVIS 2017. Due to the space limit, we provide the VFI results and their corresponding segmentation results inside the red boxes marked on the ground-truth frames. See the supplementary material for frame-by-frame comparison.

the segmentation accuracy of the baseline models by 1.1%, 0.8%, and 0.8% for AdaCoF, CDFI, and IFRNet for DAVIS 2016, respectively, and 0.9%, 0.6%, and 0.8% for AdaCoF, CDFI, and IFRNet for DAVIS 2017, respectively, in terms of $J\&F$. Although the PSNR is not correlated with the human perceptual quality of interpolated frames [42], the proposed VOS-VFI shows consistent improvements in the PSNRs of the foreground objects, indicating its effectiveness on foreground synthesis. As shown in Fig. 3, foreground moving objects are more precisely interpolated when VOS-VFI is applied to the baseline models.

Table 2 further shows the results on the widely used VFI datasets, including Vimeo90K [71], UCF101 [57]], and Xiph [43]. Note that the segmentation accuracy cannot be measured on these datasets due to the lack of segmentation annotations. Instead of the PSNR, we report other image quality metrics that are known to be highly correlated with human judgments of visual quality, including NIQE [1], PI [59], and NIMA [58]. On these advanced quality metrics, the proposed VOS-VFI provides consistent performance improvements to the baseline models.

We also used Amazon Mechanical Turk (MTurk) to collect human judgments on the visual quality of the VFI results. Specifically, each video sequence of the DAIVS 2016 validation set was frame-interpolated with a factor of 2. The MTurk users were presented with two videos obtained w/wo applying the proposed VOS-VFI training framework and conducted pairwise comparison. The left/right positions of the videos were randomized to avoid bias, and the users were asked to select one video with better visual quality.

Model	VOT2018		OTB2015	
	AUC$_\uparrow$	Robustness$_\downarrow$	AUC$_\uparrow$	Precision$_\uparrow$
AdaCoF [35]	59.2	0.440	66.8	0.876
AdaCoF-VOS	**59.8**	**0.426**	**67.4**	**0.883**
CDFI [15]	58.7	0.403	66.4	0.867
CDFI-VOS	**60.1**	**0.379**	**67.0**	**0.877**
IFRNet [32]	59.5	0.435	65.8	0.860
IFRNet-VOS	**60.3**	**0.421**	**66.5**	**0.875**
All frames	60.4	0.243	68.9	0.895

Table 4: Quantitative results of the three baseline models trained with and without VOS-VFI. The tracking performance is evaluated in terms of AUC, robustness, and precision on the VOT2018 and OTB2015 datasets.

Model	Linemod	
	REP-20px$_\uparrow$	ADD-0.1d$_\uparrow$
AdaCoF [35]	62.33	50.90
AdaCoF-VOS	**63.43**	**51.06**
CDFI [15]	63.24	50.99
CDFI-VOS	**64.09**	**51.36**
IFRNet [32]	64.16	51.60
IFRNet-VOS	**64.28**	**51.67**
All frames	99.96	96.18

Table 5: Quantitative results of the three baseline models trained with and without VOS-VFI. The 6D object pose estimation performance is evaluated in terms of RED-20px and ADD-0.1d on the Linemod dataset.

From 33 participated subjects, the proposed VOS-VFI acquired 63%, 57%, and 59% of the choice over the AdaCoF, CDFI, and IFRNet baselines, respectively.

To evaluate the performance of VOS-VFI on a large scale factor, we performed $2\times$ VFI twice to obtain $4\times$ VFI results, which is a common strategy [48, 25]. We used the HD 544p dataset [3] for the performance evaluation of $4\times$ VFI, which contains many dynamic scenes. Specifically, we extracted every fourth frame of the video sequences from the HD dataset and interpolated every intermediate three frames. As shown in Table 3, the proposed VOS-VFI improved all measured metrics. More results on other datasets can be found in the supplementary material.

4.2.2 Video Object Tracking

To show the effectiveness of VOS-VFI on the video object tracking task, we skipped every even-numbered frame in the datasets, including VOT2018 [33] and OTB2015 [66], and conducted VFI. VOT2018 contains 60 video sequences with several challenging scenarios, including fast motion and occlusion, and OTB2015 contains 100 commonly used video sequences for the performance evaluation of visual trackers. According to the benchmarks, the performance on VOT2018 was evaluated using the accuracy (average overlap over successful frames) and robustness (failure rate), and the performance on OTB2015 was evaluated using the accuracy and precision. The SiamRPN [36] was used to perform video object tracking. Table 4 shows that the video object tracking performance can be improved by applying VOS-VFI to the baseline models. In addition, the AUC performance of IFRNet-VOS is even comparable to the results obtained using all frames without VFI on VOT2018, indicating its significant effectiveness. Fig. 4 shows some video object tracking results. As can be seen, precise rendering of moving objects by VOS-VFI leads to the successful tracking of fast moving objects in the scene.

4.2.3 6D Object Pose Estimation

Last, we tested VOS-VFI on object pose estimation by skipping every even-numbered frame in the Linemod dataset [23] and interpolating the skipped frames. Linemod provides RGB-D frames containing 13 objects captured in cluttered scenes, where only one object is annotated for each sequence. The EPro-PnP model [6] was tested on the interpolated frames for 6D object pose estimation from RGB frames. The performance was evaluated in terms of REP-20px [4] and ADD-0.1d [23].

Table 5 shows that the pose estimation performance of the baseline models can be slightly improved by VOS-VFI. However, a significant performance gap exists from the result obtained using full frames. Due to a large displacement of the camera between frames and a very close distance between the camera and objects, VFI performed unsatisfactorily on Linemod. This limitation is expected to be resolved through developing VFI models robust to large camera and object motions, which is left for our future study.

4.3. Ablation Study

Objective function. To demonstrate the effectiveness of VOS-VFI, we first conducted ablation studies on the objective function by analyzing the effect of the loss terms in (1). For this experiment, CDFI was used as a baseline model. The first row of Table 6 lists the performance of the baseline model. By including the segmentation loss terms, *i.e.*, L_{segF} and L_{segB}, the segmentation accuracy was improved by 0.3% in $J\&F$. The model trained with the bi-directional consistency loss L_{segF} led to 0.6% improvement in $J\&F$. Last, the model trained with full loss terms resulted in the greatest improvement of 0.8% in $J\&F$. Consistent results can be found from the video object tracking task. Specifically, the tracking performance was improved by 1.1% (VOT2018) and 0.4% (OTB2015) in AUC when L_{segF} and L_{segB} were included. The model trained with L_{segBi} led

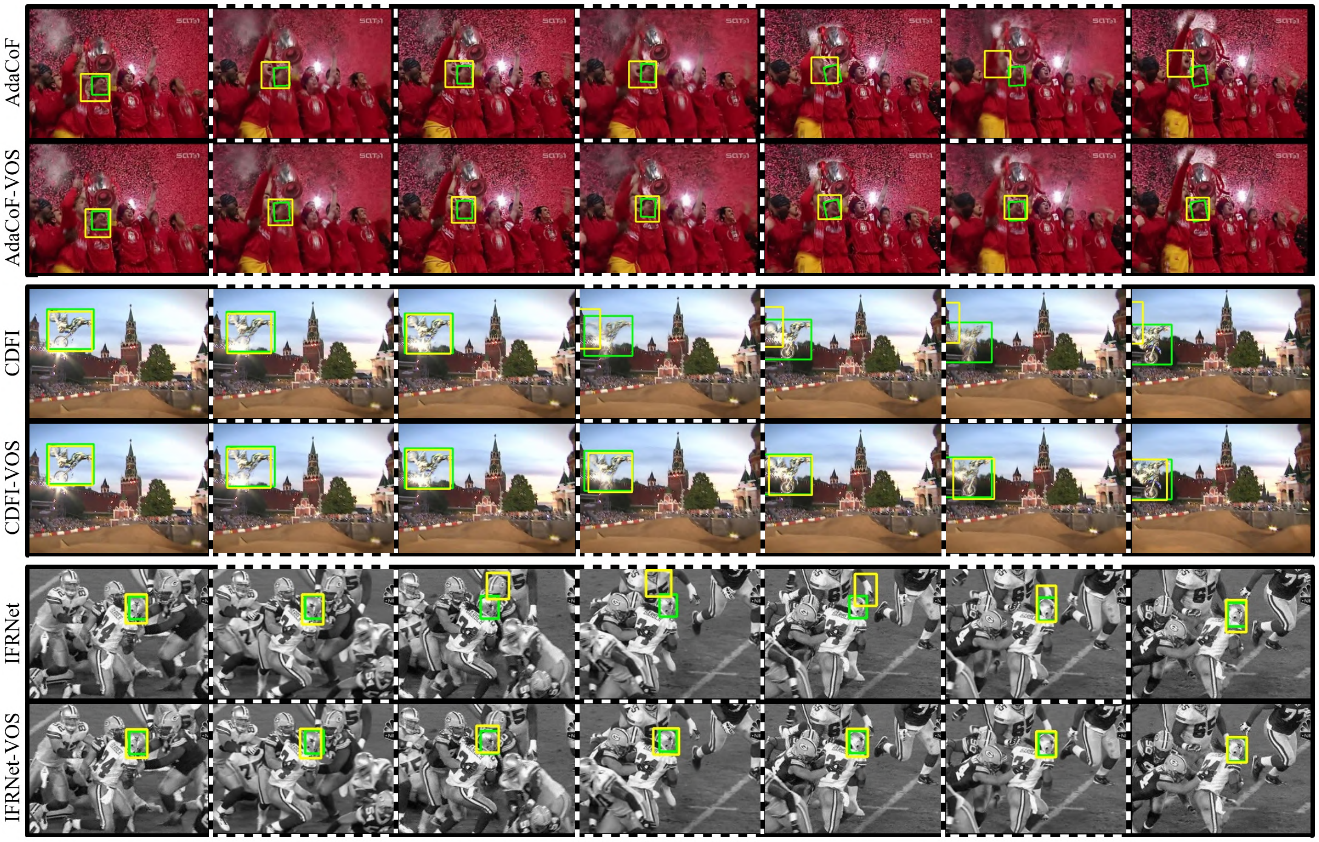

Figure 4: Examples of the visual object tracking results on the VOT2018 dataset. The odd-numbered columns and even-numbered columns (in dotted boarder lines) represent the existing and interpolated frames, respectively. The green and yellow boxes represent the ground-truth and predicted bounding boxes of objects, respectively.

Objective function			DAVIS2016				VOT2018		OTB2015	
L_{VFI}	$L_{segF,segB}$	L_{segBi}	NIQE$_{\downarrow}$	PI$_{\downarrow}$	NIMA$_{\uparrow}$	$J\&F_{\uparrow}$	AUC$_{\uparrow}$	Robustness$_{\downarrow}$	AUC$_{\uparrow}$	Precision$_{\uparrow}$
✓			3.076	2.795	3.743	86.9	58.7	0.403	66.4	0.867
✓	✓		3.082	2.793	3.736	87.2	59.8	0.397	66.8	0.868
✓		✓	3.067	2.780	3.724	87.5	59.1	0.382	66.7	0.872
✓	✓	✓	**3.054**	**2.752**	**3.783**	**87.7**	**60.1**	**0.379**	**67.0**	**0.877**

Table 6: Ablation studies on the loss functions of VOS-VFI.

to 0.4% (VOT2018) and 0.3% (OTB2015) improvements in AUC. Last, the model trained with full loss terms resulted in the greatest improvements of 1.4% (VOT2018) and 0.6% (OTB2015) in AUC.

VOS-aware module Next, to demonstrate the generalization ability of VOS-VFI, we applied the training framework using only a frame-by-frame segmentation module, as shown in Fig. 5, and the other two self-supervised VOS models [34, 60]. Specifically, in the frame-by-frame segmentation module, the object segmentation network described in Sec. 3.1 was applied to the interpolated frame and its corresponding ground-truth frame individually, resulting in $\hat{M}_t$ and M_t. As shown in Table 7, the VOS module is essential to help the VFI model synthesize moving objects. Furthermore, VOS-VFI showed consistent improvements for the three VOS models [8, 34, 60], indicating its robustness to the choice of VOS models.

Evaluation on temporal coherence To demonstrate the effectiveness of the proposed bi-directional consistency loss L_{segBi}, we evaluated the temporal coherence of VFI by

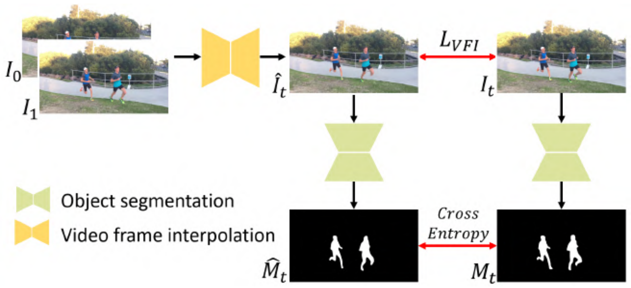

Figure 5: A training framework for the ablation study. The object segmentation model is applied to the ground-truth and interpolated frames, respectively, and the cross-entropy loss between the segmentation maps is used as an auxiliary loss term.

Model	$J\&F_{\uparrow}$	$J_{\uparrow}$	$F_{\uparrow}$
AdaCoF	85.9	84.9	86.8
AdaCoF-VOS† w/ [7]	86.0	85.0	86.9
AdaCoF-VOS w/ [8]	**87.0**	**85.9**	**88.2**
AdaCoF-VOS w/ [34]	86.9	85.8	88.0
AdaCoF-VOS w/ [60]	86.8	85.8	87.8
CDFI	86.9	85.6	88.3
CDFI-VOS† w/ [7]	86.8	85.7	88.0
CDFI-VOS w/ [8]	**87.7**	**86.4**	**88.9**
CDFI-VOS w/ [34]	87.4	86.2	88.5
CDFI-VOS w/ [60]	**87.7**	**86.4**	**88.9**

Table 7: Performance evaluation of the AdaCoF and CDFI models trained with the proposed VOS-VFI using the different pretrained VOS models for DAVIS 2016. † represents a model trained using only a frame-by-frame segmentation module.

measuring the per-frame PSNR for the interpolated frames obtained w/wo applying the bi-directional consistency loss for 4× VFI. Figure 6 shows that the bi-directional consistency loss led to temporal coherent interpolation results.

5. Conclusion

In this paper, we proposed a VOS-aware VFI training framework called VOS-VFI for interpolating intermediate frames with clear object boundaries. Using the pseudo-labels for segmentation generated from the existing VFI dataset, VOS-VFI performs VOS as an auxiliary task during training a VFI model to support additional loss functions. The additional VOS-aware loss functions contribute to interpolating objects with accurate boundaries for many challenging scenes. In particular, VOS-VFI can be applied to any VFI model during the training stage; thus, it does not require additional parameters or increase inference time. Extensive experiments with state-of-the-art VFI models on multiple benchmark datasets demonstrate that VOS-VFI can boost the performance of many vision tasks, including video object tracking and object pose estimation.

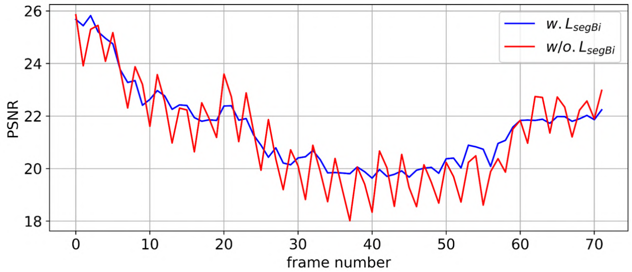

Figure 6: Example of the per-frame PSNR of the test sequence 'parkour' from DAVIS 2016 for 4× VFI.

Acknowlegement

This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C2002810).

References

[1] R. Soundararajan A. Mittal and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 6

[2] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 1, 2

[3] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):933–948, 2019. 2, 5, 7

[4] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3364–3372, 2016. 7

[5] Erkut Erdem Cagdas Bak, Aysun Kocak and Aykut Erdem. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 20(7):1688–1698, 2018. 3

[6] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2781–2790, 2022. 7

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 3, 5, 9

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 11781–11794, 2021. 2, 4, 5, 8, 9

[9] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7415–7424, 2018. 2

[10] Xianhang Cheng and Zhenzhong Chen. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[11] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020. 1, 2

[12] Yuwen He Ci Wang, Lei Zhang and Yap-Peng Tan. Frame rate up-conversion using trilateral filtering. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(6):886–893, 2010. 2

[13] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 764–773, 2017. 2

[14] Duolikun Danier, Fan Zhang, and David Bull. ST-MFNet: A spatio-temporal multi-flow network for frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3521–3531, 2022. 1, 2

[15] Tianyu Ding, Luming Liang, Zhihui Zhu, and Ilya Zharkov. CDFI: Compression-driven network design for frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8001–8011, 2021. 1, 2, 4, 5, 7

[16] Karl M Fant. A nonaliasing, real-time spatial transform technique. *IEEE Computer Graphics and Applications*, 6(1):71–80, 1986. 2

[17] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deep stereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016. 1

[18] Mingqi Gao, Feng Zheng, James JQ Yu, Caifeng Shan, Guiguang Ding, and Jungong Han. Deep learning for video object segmentation: a review. *Artificial Intelligence Review*, pages 1–75, 2022. 2

[19] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware CNN model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1142, 2015. 3

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[21] Shurui Gui, Chaoyue Wang, Qihua Chen, and Dacheng Tao. FeatureFlow: Robust video interpolation via structure-to-texture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14004–14013, 2020. 2

[22] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5038–5047, 2017. 3

[23] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In *Proceedings of the Asian Conference on Computer Vision*, pages 548–562, 2012. 7

[24] Alain Horé and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *Proceedings of the International Conference on Pattern Recognition*, pages 2366–2369, 2010. 2

[25] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision*, pages 624–642, 2022. 7

[26] Seong-Gyun Jeong, Chul Lee, and Chang-Su Kim. Motion-compensated frame interpolation based on multihypothesis motion estimation and texture optimization. *IEEE Transactions on Image Processing*, 22(11):4497–4509, 2013. 2

[27] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 1, 2

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016. 3

[29] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*, 35(6), 2016. 1

[30] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. FLAVR: Flow-agnostic video representations for fast frame interpolation. *arXiv preprint arXiv:2012.08512*, 2020. 1

[31] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. FISR: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11278–11286, 2020. 1

[32] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang.

IFRNet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022. 1, 2, 4, 5, 7

[33] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 0–0, 2018. 7

[34] Zihang Lai, Erika Lu, and Weidi Xie. MAST: A memory-augmented self-supervised tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8, 9

[35] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. AdaCoF: Adaptive collaboration of flows for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5316–5325, 2020. 1, 2, 4, 5, 7

[36] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 7

[37] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European Conference on Computer Vision*, pages 207–223, 2018. 2

[38] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 5

[40] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3532–3542, 2022. 1, 2

[41] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. UnOVOST: Unsupervised offline video object segmentation and tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2020. 2

[42] Hui Men, Hanhe Lin, Vlad Hosu, Daniel Maurer, Andres Bruhn, and Dietmar Saupe. Technical report on visual quality assessment for frame interpolation. *arXiv preprint arXiv:1901.05362*, 2019. 6

[43] Christopher Montgomery. Xiph.org video test media (derf's collection). In *Online, https://media.xiph.org/video/derf/*, 1994. 6

[44] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2018. 1, 2

[45] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017. 2

[46] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 261–270, 2017. 1, 2

[47] 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. 2

[48] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Proceedings of the European Conference on Computer Vision*, pages 109–125, 2020. 7

[49] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14539–14548, 2021. 2

[50] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 2, 3, 5

[51] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3, 5

[52] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020. 5

[53] Huazhu Fu Ming-Ming Cheng Weisi Lin Runmin Cong, Jianjun Lei and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10):2941–2959, 2019. 2, 3

[54] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen. Video frame interpolation via generalized deformable convolution. *IEEE Transactions on Multimedia*, 24:426–439, 2021. 2

[55] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17482–17491, 2022. 1, 2

[56] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: Extreme video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14489–14498, 2021. 1, 2

[57] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2, 6

[58] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 6

[59] Y. Blau *et al.* The 2018 prim challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision Workshops*, 2018. 6

[60] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. RVOS: End-to-end recurrent network for video object segmentation. In *CVPR*, 2019. 8, 9

[61] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 2

[62] Sicheng Wang, Bihan Wen, Junru Wu, Dacheng Tao, and Zhangyang Wang. Segmentation-aware image denoising without knowing true segmentation. *arXiv preprint arXiv:1905.08965*, 2019. 3

[63] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9236–9245, 2019. 2

[64] George Wolberg, HM Sueyllam, MA Ismail, and KM Ahmed. One-dimensional resampling with inverse and forward mapping functions. *Journal of Graphics Tools*, 5(3):11–33, 2000. 2

[65] Chao-Yuan Wu, Nayan Singhal, and Philipp Krähenbühl. Video compression through image interpolation. In *Proceedings of the European Conference on Computer Vision*, 2018. 1

[66] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2411–2418, 2013. 7

[67] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P. Allebach, and Chenliang Xu. Zooming Slow-Mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3370–3379, June 2020. 1

[68] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. MoNet: Deep motion exploitation for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1140–1148, 2018. 2

[69] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 585–601, 2018. 3

[70] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 32, 2019. 1, 2

[71] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2, 3, 5, 6

[72] Haoxian Zhang, Yang Zhao, and Ronggang Wang. A flexible recurrent residual pyramid network for video frame interpolation. In *Proceedings of the European Conference on Computer Vision*, pages 474–491, 2020. 1

[73] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 31, 2018. 4

[74] Shen Zheng, Changjie Lu, Yuxiong Wu, and Gaurav Gupta. SAPNet: Segmentation-aware progressive network for perceptual contrastive deraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 52–62, 2022. 3