TTA-EVF: Test-Time Adaptation for Event-based Video Frame Interpolation via Reliable Pixel and Sample Estimation

Hoonhee Cho, Taewoo Kim, Yuhwan Jeong, and Kuk-Jin Yoon
KAIST
{gnsgnsgml,intelpro,jeongyh98,kjyoon}@kaist.ac.kr

Abstract

Video Frame Interpolation (VFI), which aims at generating high-frame-rate videos from low-frame-rate inputs, is a highly challenging task. The emergence of bio-inspired sensors known as event cameras, which boast microsecond-level temporal resolution, has ushered in a transformative era for VFI. Nonetheless, the application of event-based VFI techniques in domains with distinct environments from the training data can be problematic. This is mainly because event camera data distribution can undergo substantial variations based on camera settings and scene conditions, presenting challenges for effective adaptation. In this paper, we propose a test-time adaptation method for event-based VFI to address the gap between the source and target domains. Our approach enables sequential learning in an online manner on the target domain, which only provides low-frame-rate videos. We present an approach that leverages confident pixels as pseudo ground-truths, enabling stable and accurate online learning from low-frame-rate videos. Furthermore, to prevent overfitting during the continuous online process where the same scene is encountered repeatedly, we propose a method of blending historical samples with current scenes. Extensive experiments validate the effectiveness of our method, both in cross-domain and continuous domain shifting setups. The code is available at https://github.com/Chohoonhee/TTA-EVF.

1. Introduction

Video Frame Interpolation (VFI) is a well-established problem in the field of computer vision, aiming to enhance the temporal resolution of videos. In recent times, deep learning-based VFI approaches [2, 3, 15, 20–22, 33, 35, 37, 38, 50, 60] have achieved remarkable performances across various benchmark datasets. However, accurate motion estimation becomes challenging in scenes with complex and non-linear motions due to the lack of information, often resulting in the generation of erroneous inter frames.

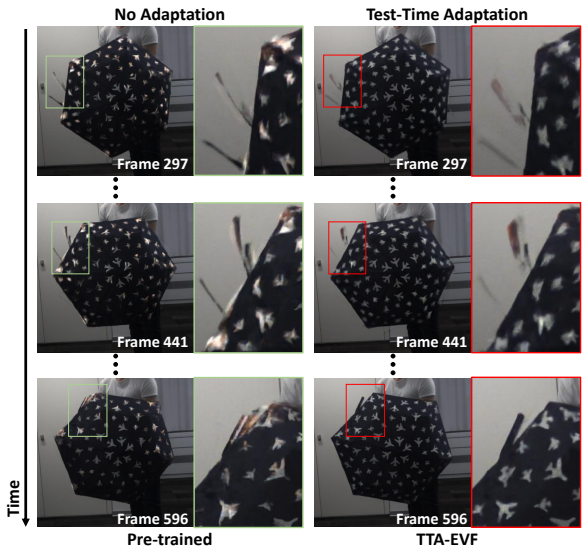

Figure 1. TTA-EVF efficiently produces high-quality results by adapting the network to the target domain in an online manner, alleviating the need for offline data supply and addressing the performance degradation observed when applying a well-trained event-based VFI network to a domain with a different distribution.

To address the inherent lack of intermediate information in frame cameras, recent researches [16, 24, 51, 52, 56] have explored the potential of event cameras as a promising solution for VFI, especially in scenarios involving complex motion. Event cameras sense dynamic changes in pixel intensity, triggering an event when the change surpasses a predefined threshold. The incorporation of event cameras as additional devices in VFI mitigates the challenges associated with modeling complex motion.

Typically, frame-based VFI methods [17, 33, 38] are evaluated on test datasets (*e.g.*, SNU-FILM [12], X4K1000FPS [43]) with distributions different from their training dataset (*e.g.*, Vimeo90K [61]). Frame-based VFI demonstrates a certain level of generalization ability even without additional modules, providing evidence that it can be applied to real scenario applications, such as dealing with changes in camera devices. These device and environment

variations are inherent in VFI, which must learn from High-Frame-Rate (HFR) cameras and perform inference on Low-Frame-Rate (LFR) cameras, making them unavoidable.

However, when training event-based VFI using a source dataset and testing on a different domain, there is a significant performance drop compared to frame-based approaches. This is attributed not only to the inherent noise from external environment (*e.g.*, illumination) but also to the distribution variations due to different camera settings (*e.g.*, event trigger threshold). To address this, one can utilize self-supervised learning [16, 39] with low-frame-rate and event data from the target domain, but there is a drawback in applying them online. In practical application, event cameras exhibit continuous changes in distribution with variations in lighting conditions. It is not feasible to acquire new data for novel scenes each time they arise and wait for the network to learn offline. In this regard, an online adaptation of the network to the environment is essential from a practical perspective and is well-known as Test-Time Adaptation (TTA) [47, 53, 63]. TTA assumes a setting that does not allow access to source data and prefers online learning for adaptation. In our case, the target domain lacks ground-truth data, which means no high-frame-rate video is available in the target domain.

In this paper, we propose Test-time Adaptation for Event-based Video Frame Interpolation (TTA-EVF) to alleviate performance degradation in the face of domain shifts. As shown in Fig. 1, when encountering a new domain without performing adaptation, even well-trained networks exhibit performance drops. In contrast, TTA-EVF reliably and rapidly fine-tunes in the target domain through self-training, even without high-frame-rate videos. For TTA in event-based VFI, there are two main challenges: Firstly, even without exposure to HFR videos in the target domain, the network should adapt to the scene while simultaneously learning knowledge for HFR generation from LFR data. To address this challenge, we propose Reliable Pixel Sampling (RPS), a self-training scheme for selecting pixels from generated HFR videos, utilizing a confidence estimation based on a teacher-student framework [49]. The second challenge is overfitting when continuously provided with data that has a similar distribution within a successive sequence within online learning. To mitigate this, we employ a memory bank that stores only reliable samples from the generated images. We then propose Patch-Mixed Sampling (PMS) by blending reliable samples with current images, effectively preventing overfitting through an augmentation approach.

The main contributions of our works are as follows: (I) We propose a novel Test-Time Adaptation (TTA) framework for event-based VFI. (II) We introduce a novel Event-RGB Distribution Shift (ERDS) dataset for event-based video frame interpolation, where the distribution of the dataset is continually changing. (III) We compare our method with existing works and demonstrate its superiority by outperforming supervised methods.

2. Related Works

Event-based Video Frame Interpolation. Frame-based VFI methods [2, 3, 15, 20–22, 33, 35, 37, 38, 50, 60] have continuously achieved performance improvements; however, it still has weaknesses in handling complex motion and requires motion approximation when synthesizing intermediate frames at arbitrary time. Event-based VFI methods [30, 41, 44, 45, 55, 59, 64, 65], which utilize event cameras as an additional sensor [6, 7, 9, 28, 34, 45], enable not only handling complex motion but also interpolating intermediate frames for irregular objects. However, there is still room for improvement compared to frame-based VFI methods. Typically, frame-based VFI methods perform cross-domain performance evaluations, considering real scenario applications. Well-trained frame-based VFI methods exhibit domain generalization ability, maintaining considerable performance even on different domains without the need for additional training process for adaptation. On the other hand, event-based VFI experiences a significant drop in performance when moving to different domains [16]. Therefore, resolving cross-domain challenges in event-based VFI is vital for widespread applications.

Test-Time Adaptation. Unlike Unsupervised Domain Adaptation [8, 48, 58], which necessitates offline learning for each domain change and relies on labeled data from the source domain, Test-Time Adaptation (TTA) is more challenging setup as it adapts to the target domain without accessing source data. In recent times, the significance of data privacy has escalated, leading to increased interest in TTA, a source-free domain adaptation approach. This is primarily due to the inefficiency of accessing source domain data during inference. TTA is actively being researched in the image domain [5, 14, 19, 27, 31, 32, 36, 40, 42, 47, 53, 63, 66], especially in areas like object recognition and segmentation. While there is an approach of its application in event-based visual recognition, Ev-TTA [23], there is still a lack of research in the domain of video frame interpolation, particularly in relation to events. The existing work, MetaVFI [11], proposed for the frame-based VFI task, learns to estimate scene-adaptive motion with minimal parameter updates through meta-learning [13]. However, this method is well-suited when there is not a significant domain gap between training and testing. We demonstrate that this approach is suboptimal when applied to event-based VFI with a substantial domain gap between the training and test sets.

We propose the first-ever Test-Time Adaptation (TTA) approach for event-based Video Frame Interpolation (VFI). This is essential because event distributions can vary significantly due to external environmental factors and camera parameter settings even when using the same camera.

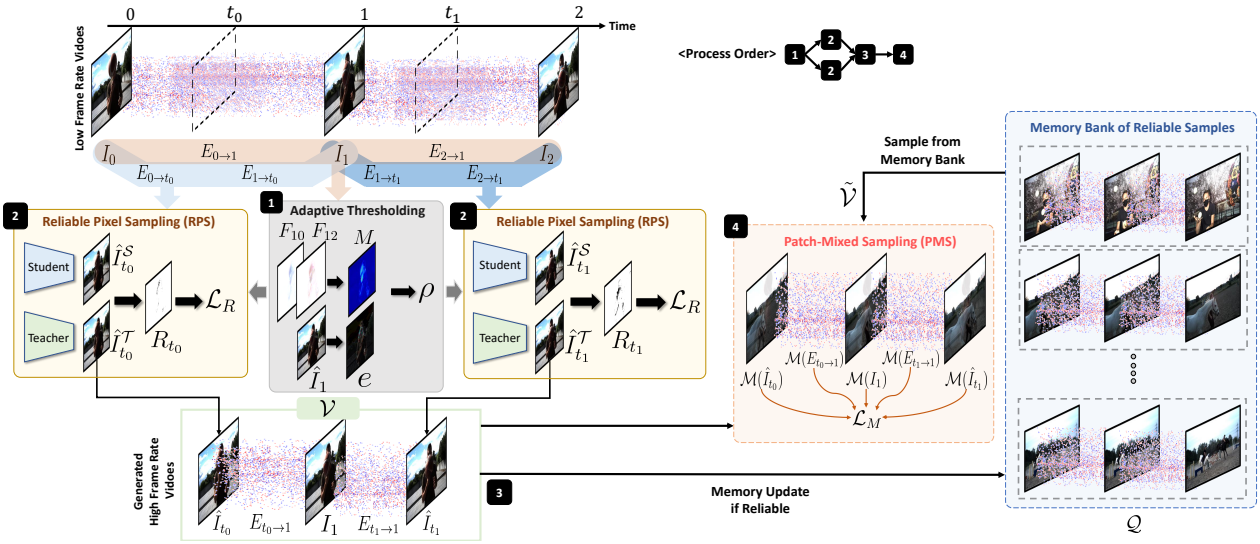

Figure 2. **Overview of our Test-Time Adaptation for Event-based Video Frame Interpolation (TTA-EVF) framework.** Our approach involves four steps for test-time training in Ev-VFI: 1) We calculate the adaptive threshold, ρ, for RPS using the low frame rate image triples, I_0, I_1, and I_2 with interval events $E_{0\to1}$ and $E_{2\to1}$. 2) Reliable Pixel Sampling (RPS) is applied for arbitrary t_0 between 0 and 1, and arbitrary t_1 between 1 and 2, and in this process, even without intermediate images for training, reliable pixels are estimated through the student-teacher framework and self-training is performed. 3) Using the estimated intermediate frames $\hat{I}_{t_0}$ and $\hat{I}_{t_1}$ along with I_1, we form a new triplet, $\mathcal{V}$, check if the entire sampled generated image is reliable, and in that case, update the memory bank. 4) Patch-Mixed Sampling (PMS) regularizes the network by mixing the generated videos, $\mathcal{V}$ with videos from the memory bank, $\tilde{\mathcal{V}}$.

Furthermore, our method follows the recent trend of TTA by addressing the adaptation to unknown distribution shifts encountered during test-time in an online manner.

3. Proposed Method

3.1. Overview

Given key frames I_0, I_1 and the inter-frame event stream $\mathcal{E}_{0\to1}$, Event-based Video Frame Interpolation (Ev-VFI) networks aim to generate the intermediate frame I_t on an arbitrary time $t \in [0, 1]$. Following the previous works, we convert a event stream as a voxel grid [67]. We denote $E_{t_a\to t_b}$ as a voxel grid form between times t_a and t_b.

TTA-EVF adapts a pre-trained Ev-VFI network trained on the source domain to a target domain with domain shift in the device and setting. In the pre-training phase, we train the network using source data with high-frame rate videos in a supervised manner (Sec. 3.2). We then utilize this pre-trained Ev-VFI network for test-time adaptation using target data without accessing the source data (Sec. 3.3).

3.2. Pre-training on Source Dataset

In general, TTA methods perform test-time training stably by updating only a subset of network parameters, such as Batch Normalization (BN) [18]. As mentioned in [53], updating the full parameters at test-time, especially in an online manner, leads to instability and makes it challenging to achieve optimal performance. However, in the case of low-level vision tasks like VFI, normalization layer (*e.g.* BN) is often avoided by most networks [20, 24, 51] since it removes range flexibility and can lead to a performance decrease [29]. Some alternatives for handling this issue include using slightly more flexible normalization techniques (*e.g.* instance normalization) or selectively updating specific convolutional layers. Updating only the parameters of the normalization layer in VFI can lead to sub-optimal performance when there is significant variance between image patches in the training and test environments. Conversely, updating specific convolution layers with more parameters can remove all statistical modulation, causing instability during the test-time training process. To address this issue, we propose a solution by designing a Norm-Residual (NR) Block that includes both convolution and normalization layers. This approach aims to strike a balance between updating parameters effectively and maintaining statistical modulation during the learning process.

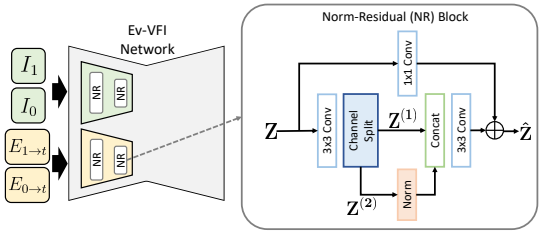

Figure 3. The architecture of Norm-Residual (NR) Block.

As shown in Fig. 3, given a feature $\mathbf{Z}$, NR Block first split a feature into two feature maps $\mathbf{Z^{(1)}}$ and $\mathbf{Z^{(2)}}$ along the channel dimension. The first map $\mathbf{Z^{(1)}}$ is skipped to keep

the context information, and rest feature $\mathbf{Z}^{(2)}$ is normalized by normalization layer. Then, we fuse the two features by concatenate and convolutional layer. Finally, we apply a element-wise summation with original feature $\mathbf{Z}$ to obtain the output $\hat{\mathbf{Z}}$. We use the group normalization [57] in normalization layer, which shows the best performance in our experiments. For efficiency, we exclusively deploy the NR Block in the encoders of the networks, replacing the conventional residual blocks of existing networks. Therefore, TTA-EVF is a general and model-agnostic method specifically crafted to be adaptable to any network architecture used in Ev-VFI. Further details can be found in the *supple*.

Network Optimization. The modified Ev-VFI network with the NR Block is trained on the source dataset using ground truth with high-frame-rate videos. Specifically, we optimize the Ev-VFI network, $\mathcal{F}$, by the supervision as:

$$\arg\min_{\theta(\mathcal{F})}(\mathcal{L}_{\mathcal{F}}(\hat{I}_t, I_t)), \tag{1}$$

where $\mathcal{L}_{\mathcal{F}}$ is the loss function proposed by each Ev-VFI network and the generated intermediate frame $\hat{I}_t$ is obtained by $\hat{I}_t = \mathcal{F}\left(I_0, I_1, E_{0\to t}, E_{1\to t}\right)$.

3.3. Test-time Adaptation on Target Dataset

In our TTA setup, the model generates intermediate frames at each iteration and simultaneously updates the parameters of the NR Block. Since the target domain only contains LFR frames, not HFR frames, self-supervised optimization using only LFR is required.

3.3.1 Reliable Pixel Sampling (RPS)

In the target domain, even without High-Frame-Rate (HFR) ground truth, when given Low-Frame-Rate (LFR) videos I_0, I_1, I_2 as triplets, and the events between them $E_{0\to1}$ and $E_{2\to1}$, the network can be trained using the following self-supervised loss:

$$\mathcal{L}_S = \mathcal{L}_{\mathcal{F}}(\hat{I}_1, I_1), \tag{2}$$

where $\hat{I}_1 = \mathcal{F}\left(I_0, I_2, E_{0\to1}, E_{2\to1}\right)$. However, this is problematic as the network cannot learn the knowledge of generating HFR because it is generating LFR again from temporally subsampled LFR images, leading to the loss of the ability to generate HFR trained from the source domain. As a way to address this, applying the cyclic consistency loss [16, 39] from the existing unsupervised VFI is one option. However, directly applying these methods to TTA of Ev-VFI is intractable, since there is no guarantee that the generated intermediate frames will be clean and artifact-free on target domains. Therefore, we need a new paradigm.

As shown in Fig. 2, given the I_0, I_1, I_2 triplet, we design Reliable Pixel Sampling (RPS) for the left intermediate time t_0 and the right intermediate time t_1. Here, t_0 and t_1 are selected at random from the time intervals 0 to 1 and 1 to 2, respectively. Since the RPS process is identical for both t_0 and t_1, we introduce it here for t_0 only. Reliable data sampling without the intermediate ground truth frames, I_{t_0}, is highly challenging and can lead to unstable learning if estimated incorrectly. To address this, we adopt the teacher-student framework [49] commonly used for pseudo labeling [10, 58] in other tasks. In this framework, the student learns from pseudo labels through self-training, while the teacher is updated solely from the momentum using the parameters of the student. The motivation behind proposed RPS stems from the observation that pixels estimated similarly by two networks with slightly different parameters, yet achieving similar performance in VFI, are likely to be reliable, whereas parts that are difficult to estimate and prone to errors would result in differing predictions between the two networks. Hence, we classify pixels consistently estimated by both teacher-student networks as reliable pixels, while considering pixels with inconsistent predictions as unreliable. Given intermediate frames generated from the student network $\mathcal{F}_{\mathcal{S}}$ and the teacher network $\mathcal{F}_{\mathcal{T}}$ denoted as $\hat{I}^{\mathcal{S}}_{t_0}$ and $\hat{I}^{\mathcal{T}}_{t_0}$ respectively, the reliable map, R_{t_0}, with spatial dimensions $H \times W \times 1$ (where H and W represent the height and width of the image I_0) is defined as follows:

$$R_{t_0}(i,j) = \begin{cases} 1\ , \text{ if } \|\hat{I}^{\mathcal{S}}_{t_0}(i,j) - \hat{I}^{\mathcal{T}}_{t_0}(i,j)\| < \rho, \\ 0\ , \text{ otherwise,} \end{cases} \tag{3}$$

where $i = 1, \ldots, H$, $j = 1, \ldots, W$, and ρ denote the threshold. When estimating such reliability, it is crucial to determine a specific threshold. However, assigning the same threshold to all scenes is inappropriate since the performance of VFI significantly varies across scenes. To achieve adaptive thresholding for each scene, we utilize the LFR frames I_0 and I_2, along with the interceding events $E_{0\to1}$ and $E_{2\to1}$, to estimate $\hat{I}_1$. Subsequently, we compute a pixel-wise error map, $e = \|\hat{I}_1 - I_1\|$, between $\hat{I}_1$ and I_1, allowing us to calculate ρ, using the average value of e.

However, this approach encounters issues when dealing with scenes that contain motionless, static regions (*e.g.*, backgrounds). In such cases, these areas exhibit minimal pixel-wise errors, which leads to an inadequate adaptive threshold calculation for the scene if these pixels are included in the average calculation. To address this, we propose an approach that excludes motionless regions based on optical flow to refine the threshold calculation. Since most Ev-VFI methods compute optical flow towards the intermediate frame from key frames, we can directly calculate a motion magnitude map [54], M, from the intermediate output of optical flow, denoted as F_{10} and F_{12}, without requir-

Algorithm 1 PMS Pseudo Code

Input: Generated HFR set $\mathcal{V} = \{\hat{I}_{t_0}, I_1, \hat{I}_{t_1}, E_{t_0 \to 1}, E_{t_1 \to 1}\}$

Initialize $\mathcal{L}_M \leftarrow 0$
Get reliability of $\mathcal{V}$ based on Eq.(9)
Sample $\tilde{\mathcal{V}} = \{\tilde{I}_{\tilde{t}_0}, \tilde{I}_1, \tilde{I}_{\tilde{t}_1}, \tilde{E}_{\tilde{t}_0 \to 1}, \tilde{E}_{\tilde{t}_1 \to 1}\}$ from $\mathcal{Q}$
Sample mixing ratio $\lambda \in (0.7, 1.0)$
if $\mathcal{V}$ is reliable **then**
    $\mathcal{M}_{t_0} = \lambda \hat{I}_{t_0} + (1-\lambda)\tilde{I}_1$, $\mathcal{M}_{t_1} = \lambda \hat{I}_{t_1} + (1-\lambda)\tilde{I}_1$
    $\mathcal{M}_1 = \lambda I_1 + (1-\lambda)\tilde{I}_1$
    $\mathcal{D}_{t_0 \to 1} = E_{t_0 \to 1}$, $\mathcal{D}_{t_1 \to 1} = E_{t_1 \to 1}$
    Push $\mathcal{V}$ into memory bank $\mathcal{Q}$
else
    $\mathcal{M}_{t_0} = (1-\lambda) I_1 + \lambda \tilde{I}_{\tilde{t}_0}$, $\mathcal{M}_{t_1} = (1-\lambda) I_1 + \lambda \tilde{I}_{\tilde{t}_1}$
    $\mathcal{M}_1 = (1-\lambda) I_1 + \lambda \tilde{I}_1$
    $\mathcal{D}_{t_0 \to 1} = \tilde{E}_{\tilde{t}_0 \to 1}$, $\mathcal{D}_{t_1 \to 1} = \tilde{E}_{\tilde{t}_1 \to 1}$
if len($\mathcal{Q}$) $>$ max queue size **then**
    Pop oldest ones out of $\mathcal{Q}$
$\mathcal{L}_M \leftarrow \mathcal{L}_{\mathcal{F}}(\mathcal{M}_1, \mathcal{F}_{\mathcal{S}}(\mathcal{M}_{t_0}, \mathcal{M}_{t_1}, \mathcal{D}_{t_0 \to 1}, \mathcal{D}_{t_1 \to 1}))$

Output: $\mathcal{L}_M$

ing additional computations as:

$$M(i,j) = (\|F_{10}(i,j)\| + \|F_{12}(i,j)\|)/2, \tag{4}$$

where $\|\cdot\|$ denote the magnitude of the motion vector. The motion mask $\widehat{M}$, which excludes motionless areas based on the motion magnitude map for each pixel, can be created as follows:

$$\widehat{M}(i,j) = \begin{cases} 1 & \text{if } M(i,j) \geq m, \forall i, \forall j \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where, m represents a predefined minimal amount of motion, and in our experiment, we use a value of 1. Then, we calculate the scene adaptive threshold as:

$$\rho = \frac{1}{|\widehat{M}|} \sum \widehat{M} \odot e. \tag{6}$$

We define the self-training loss between the teacher and student networks using a pixel-wise reliable map in Eq.(3) and Eq.(6):

$$\mathcal{L}_R^{t_0} = \|\hat{I}_{t_0}^{\mathcal{S}}(i,j) - \hat{I}_{t_0}^{\mathcal{T}}(i,j)\| \odot R_{t_0}, \tag{7}$$

Finally, the reliable pixel-wise loss, taking into account both intermediate times t_0 and t_1, is defined as follows:

$$\mathcal{L}_R = \mathcal{L}_R^{t_0} + \mathcal{L}_R^{t_1}, \tag{8}$$

3.3.2 Patch-Mixed Sampling (PMS)

The issues that can arise in TTA include overfitting to specific scenes due to the sequential data for the same scene being continuously fed, as well as the problem of catastrophic forgetting [4, 26] where existing knowledge is lost.

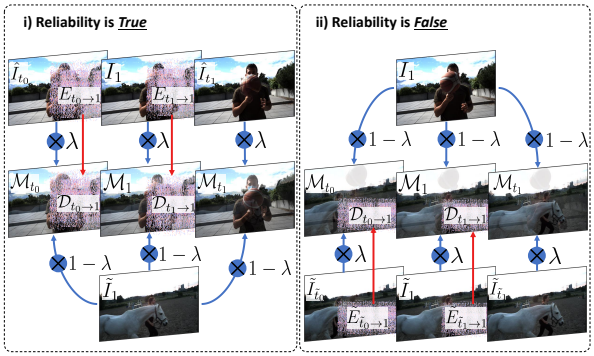

Figure 4. The examples of patch-mixed sampling (PMS) process.

To address this, we propose the Patch-Mixed Sampling (PMS) method. In order to avoid overfitting to specific scenes, PMS involves sampling from the memory bank, which accumulates reliable samples from the generated videos of each iteration. The detailed process of PMS is provided in Algorithm 1.

Reliability Check. Denoting the HFR video and event set, including $\hat{I}_{t_0}$ and $\hat{I}_{t_1}$, from the RPS process as $\mathcal{V}$, we assess the reliability for the generated frames. If the pixel-wise reliability map R obtained from the RPS module exceeds η fraction of the resolution of the entire image, we deem the generated images as reliable. This is outlined as follows:

$$\text{Reliablity} = \begin{cases} \text{True, if } \frac{\|R_{t_0}\|}{H \times W} > \eta \text{ and } \frac{\|R_{t_1}\|}{H \times W} > \eta \\ \text{False, otherwise,} \end{cases} \tag{9}$$

Patch Mixing. Figure 4 shows an example of a patch mixing process. If the currently generated HFR video set $\mathcal{V}$ is reliable (left of Fig. 4), we fetch a single intermediate frame $\tilde{I}_1$ from the memory bank and blend it. Since the mixed frames, M_{t_0}, M_1, and M_{t_1}, all incorporate the same $\tilde{I}_1$, any events that require pixel changes can't arise from $\tilde{I}_1$, thus we directly retain the events $E_{t_0 \to 1}$ and $E_{t_1 \to 1}$ from $\mathcal{V}$. On the other hand, if $\mathcal{V}$ is unreliable (right of Fig. 4), we retrieve both images and event sets from the memory bank. From the currently generated $\mathcal{V}$, we only fetch I_1 and blend it. Similar to before, since the same I_1 is blended into all three images, events originating from I_1 do not occur. Therefore, we directly retain the memory bank events $\tilde{E}_{\tilde{t}_0 \to 1}$ and $\tilde{E}_{\tilde{t}_1 \to 1}$. For the mixing ratio of blending images, we adopt a beta distribution [62] that generates values between 0.7 and 1.0. We assign a larger ratio to the reliable set among the two sets being blended. Finally, using the patch-mixed sets, we train the student network of the RPS using the loss of Ev-VFI as follows:

$$\mathcal{L}_M = \mathcal{L}_{\mathcal{F}}(\mathcal{M}_1, \mathcal{F}_{\mathcal{S}}(\mathcal{M}_{t_0}, \mathcal{M}_{t_1}, \mathcal{D}_{t_0 \to 1}, \mathcal{D}_{t_1 \to 1})). \tag{10}$$

In summary, the total objective $\mathcal{L}$ of TTA-EVF is as follows:

$$\mathcal{L} = \mathcal{L}_S + \lambda_1 \mathcal{L}_R + \lambda_2 \mathcal{L}_M \tag{11}$$

where λ_1, λ_2 are the hyper-parameters.

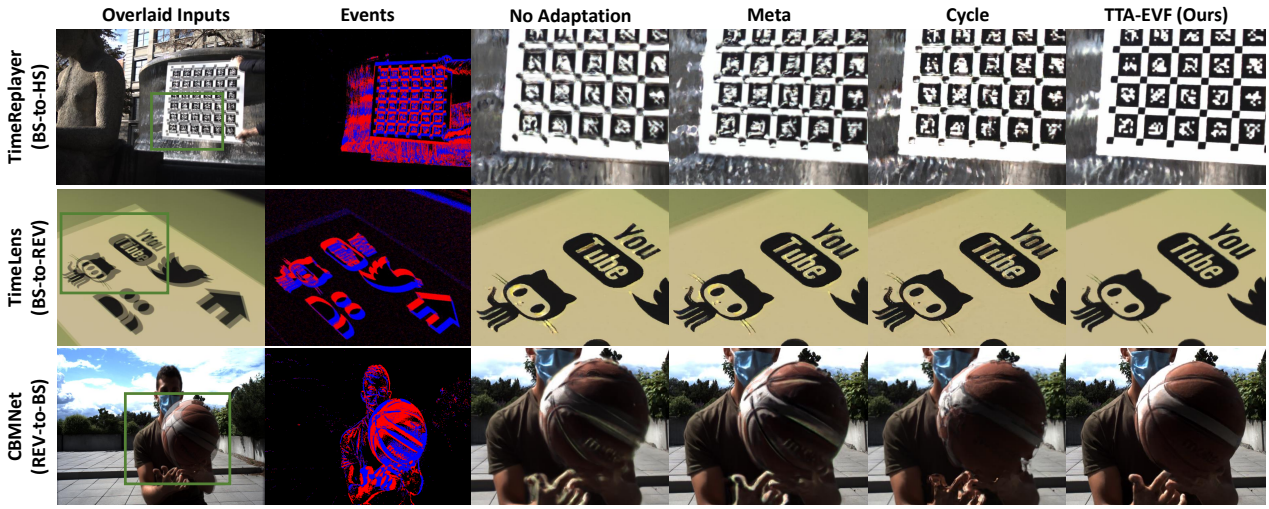

Figure 5. Visual comparisons of interpolated frames on multiple cross-domain settings using various models.

4. Experiments

4.1. Implementation Details

All networks are trained using an Adam optimizer [25] with a batch size of 2 and a learning rate of 1×10^{-4}, for both the source and target domains. For efficiency, during online training in the target domain, we crop images to a resolution of 320×320. In our approach, we perform network updates only once for each triplet provided. For instance, when interpolating 7 skipped frames, we generate a total of 14 frames from I_0, I_1, and I_2, and update the network parameters only once using the loss from Eq.(11), making it highly efficient. All experiments are conducted on a single NVIDIA TITAN RTX. We set the η as 0.9 in Eq.(9), bin size of voxel grid as 5, and λ_1, λ_2 as 1, 0.5 in Eq.(11).

4.2. Test-Time Adaptation Results

During the pre-training phase, we utilize the train sequences from the source domain for learning, and in the target domain, we follow the same sequential online approach as in the traditional TTA setup [53], using only the test sequences without utilizing the train set.

Our framework can be applied to existing Ev-VFI methods regardless of the model. We conduct experiments by integrating it with TimeLens [51] and CBMNet [24], using publicly available code provided by the authors. For TimeReplayer [16], as public code is unavailable, we re-implement it based on the foundational code [20] as described in the paper. To the best of our knowledge, we are the first to propose test-time adaptation for event-based video frame interpolation. Therefore, for comparison, we adopt the previously proposed meta-learning approach [11] for frame-based VFI and the self-supervised optimization [16] for Ev-VFI. When training at test time, the performance tends to be low if we simply apply the existing methods [16] as is. Therefore, we keep our NR Block update and TTA setup intact while only incorporating the cyclic self-supervised mechanism [16] into Eq.(11).

Table 1. Quantitative comparisons with existing method in cross-domain datasets. S and T denote the source and target datasets.

S	T	Methods	TimeReplayer [16] PSNR	TimeReplayer [16] SSIM	TimeLens [51] PSNR	TimeLens [51] SSIM	CBMNet [24] PSNR	CBMNet [24] SSIM
B	R	No Adapt	32.24	0.9154	32.04	0.9073	31.98	0.9081
		Meta [11]	32.75	0.9180	32.43	0.9112	32.21	0.9109
		Cycle [16]	32.87	0.9208	33.01	0.9165	33.09	0.9214
		TTA-EVF	**34.07**	**0.9286**	**34.24**	**0.9299**	**34.42**	**0.9312**
B	H	No Adapt	29.78	0.8378	29.96	0.8413	29.87	0.8395
		Meta [11]	30.13	0.8393	30.14	0.8341	30.22	0.8407
		Cycle [16]	30.32	0.8104	30.66	0.8441	30.75	0.8466
		TTA-EVF	**31.45**	**0.8551**	**32.13**	**0.8599**	**32.07**	**0.8584**
R	B	No Adapt	21.94	0.6434	22.32	0.6745	21.89	0.6427
		Meta [11]	22.83	0.6818	23.14	0.7020	22.75	0.6822
		Cycle [16]	23.98	0.7154	24.11	0.7223	24.43	0.7305
		TTA-EVF	**25.85**	**0.7638**	**26.91**	**0.7752**	**27.16**	**0.7768**
R	H	No Adapt	27.83	0.8158	27.61	0.8121	27.92	0.8222
		Meta [11]	28.11	0.8206	28.32	0.8234	28.54	0.8299
		Cycle [16]	28.43	0.8306	29.12	0.8389	29.43	0.8410
		TTA-EVF	**30.33**	**0.8464**	**31.83**	**0.8587**	**31.86**	**0.8590**

TTA in Cross-domain Datasets. We conduct cross-domain experiments using three existing high-quality real event datasets: BS-ERGB [52], HS-ERGB [51], and High-REV [46]. We refer to each datasets as B, H, and R, respectively. In each experiment, we alternate between the source and target datasets, but HS-ERGB exclusively provides with test sequences and is used solely as the target dataset.

Table 1 presents the quantitative results in cross-domain datasets. We can confirm that TTA-EVF has significant advantages over other methods. First, Meta [11] using the meta-learning shows an overall but subtle improvement in performance. It indicates that it is not particularly effective in adapting to entirely new domains, especially when the distribution of event data changes, operating independently for each scene. In contrast, we continuously adapt online to scenes, learning in an optimally directional manner for the entire domain rather than for each individual scene. Further-

Table 2. Quantitative comparisons with existing method in continuous domain shifting datasets.

Networks	Methods	Target Sequences													
		1		2		3		4		5		6		Total	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
TimeReplayer [16]	No Adapt	34.67	0.9090	28.52	0.6707	33.20	0.7480	30.94	0.7024	34.87	0.9069	36.50	0.9054	32.81	0.7960
	Meta [11]	33.93	0.9063	28.65	0.6732	32.34	0.7303	31.14	0.7045	34.91	0.9083	36.95	0.9106	32.91	0.7996
	Cycle [16]	34.74	0.9098	27.43	0.6674	32.54	0.7345	31.22	0.7051	35.10	0.9091	37.05	0.9097	32.99	0.7968
	TTA-EVF	**35.05**	**0.9101**	**29.44**	**0.6765**	**33.47**	**0.7529**	**31.90**	**0.7175**	**36.79**	**0.9213**	**37.27**	**0.9144**	**33.71**	**0.8040**
TimeLens [51]	No Adapt	33.91	0.8955	28.17	0.6664	32.71	0.7428	30.58	0.7024	33.99	0.9002	34.29	0.8924	32.07	0.7896
	Meta [11]	34.05	0.9001	27.84	0.6642	32.80	0.7412	30.15	0.7022	33.69	0.8907	34.49	0.9004	32.09	0.7898
	Cycle [16]	34.01	0.9008	27.96	0.6680	32.41	0.7435	30.22	0.7041	34.07	0.9008	34.98	0.9013	32.14	0.7906
	TTA-EVF	**34.41**	**0.9055**	**28.63**	**0.6737**	**33.09**	**0.7493**	**31.29**	**0.7100**	**34.94**	**0.9059**	**36.05**	**0.9040**	**32.80**	**0.7974**
CBMNet [24]	No Adapt	32.70	0.9014	27.19	0.6597	32.05	0.7326	30.14	0.7015	32.24	0.8962	36.82	0.9117	31.42	0.7879
	Meta [11]	32.84	0.9026	27.32	0.6608	32.14	0.7334	30.16	0.7012	32.31	0.8974	36.65	0.9101	31.48	0.7885
	Cycle [16]	33.62	0.9097	27.82	0.6695	32.61	0.7372	30.34	0.7058	32.85	0.9013	37.19	0.9145	32.41	0.7988
	TTA-EVF	**35.82**	**0.9137**	**29.25**	**0.6809**	**33.26**	**0.7540**	**31.65**	**0.7203**	**35.12**	**0.9104**	**37.54**	**0.9288**	**33.53**	**0.8023**

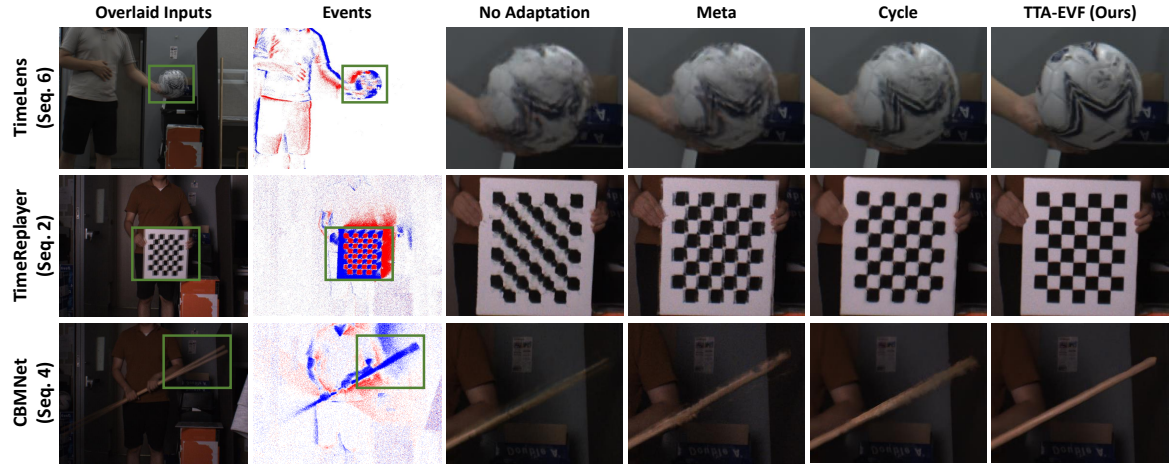

Figure 6. Visual comparisons of interpolated frames on continuous domain shifting settings. Please zoom for better view.

Table 3. Comparison results between our TTA method and existing supervised/unsupervised methods in the target domain. * denotes that the values are taken from the original paper.

S	T	Methods	Online	TimeReplayer [16]		TimeLens [51]	
				PSNR	SSIM	PSNR	SSIM
R	R	Supervised		-	-	32.81*	0.901*
		Unsupervised		32.42	0.898	-	-
B	R	TTA-EVF	✓	**34.07**	**0.929**	**34.24**	**0.930**
H	H-far	Supervised		-	-	32.31*	0.869*
		Unsupervised		30.07*	0.834*	-	-
B	H-far	TTA-EVF	✓	**31.72**	**0.859**	**32.42**	**0.870**
H	H-close	Supervised		-	-	31.68*	0.835*
		Unsupervised		29.83*	0.816*	-	-
B	H-close	TTA-EVF	✓	**30.95**	**0.829**	**32.11**	**0.860**

more, when we examine the results of Cycle [16], which incorporate a self-supervised loss into our TTA approach, it becomes evident that simply maintaining continuous online updates does not yield significant performance gains. In the case of Cycle [16], learning cyclically with generated videos during limited online updates is unable to amplify adaptation ability in situations where unavoidable errors occur as domains change. In contrast, our method identifies significant errors in pixel and samples for each scene, producing substantial improvements even within the same setting. Figure 5 also demonstrates the effectiveness of our method across various datasets and networks.

In addition, we present the comparison of our TTA approaches with supervised/unsupervised methods in Table 3. Despite the fact that our method involves online updates on the target dataset after pre-trained from other source domain datasets, it outperforms methods trained on the target dataset using supervised or unsupervised approaches. For instance, on the HighREV (R) dataset using the TimeLens [51] model, our approach achieved a 1.43 dB higher PSNR than supervised learning, even when trained in a different domain, highlighting its ability to bridge domain gaps and improve the versatility of event-based VFI.

TTA in Continuous Domain Shift. We experiments in a highly challenging setting where the distribution of the dataset continually undergoes significant changes. Since there were no existing datasets with consistently changing domains in high-quality RGB-Event data, we acquire the novel Event-RGB Distribution Shift (ERDS) dataset. In the ERDS dataset, the training dataset includes HFR videos with event data in typical lighting conditions. The test dataset, on the other hand, includes scenarios where lighting conditions are consistently altered, and the event trigger threshold is adjusted accordingly to obtain sufficient information from the events. We present the results in Table 2. In the highly challenging setting of continuous domain shift-

Table 4. Ablation study of the proposed components.

$\mathcal{L}_S$	NR Block	$\mathcal{L}_R$	$\mathcal{L}_M$	Cross	Continuous
No Adaptation				29.96	31.42
✓				30.42	31.98
✓	✓			31.53	32.76
✓	✓	✓		31.84	33.01
✓	✓		✓	31.92	33.15
✓	✓	✓	✓	**32.13**	**33.53**

Table 5. Results based on various parameter update method.

Update	PSNR ↑	SSIM ↑
Full Parameters	31.04	0.8489
Encoder Parameters	30.89	0.8455
Batch Normalization (BN) [18]	30.44	0.8423
Layer Normalization (LN) [1]	30.92	0.8477
NR Block w/. BN [18]	31.57	0.8502
NR Block w/. LN [1]	31.94	0.8524
NR Block w/. GN [57]	**32.13**	**0.8599**

ing, other methods fail to deliver significant improvements. In contrast, our method can achieve substantial performance gains compared to competing methods. Figure 6 clearly demonstrates the strengths of our method.

5. Ablation Study

In this section, we conduct experiments to access the efficacy of the proposed approaches. In the cross-dataset setting, we use the TimeLens [51] model, training it with BS-ERGB and evaluating it on HS-ERGB. In the continuous setting, we utilize the CBMNet [24] model.

Effectiveness of each components. We conduct experiments in Tab. 4 while adding each component of TTA-EVF one by one. When optimizing the network using only the loss $\mathcal{L}_S$ in Eq.(2) under the TTA setup, it can be observed that there is not a significant improvement in performance. This is because directly self-training with only LFR videos prevents the network from learning the knowledge of generating HFR, and as training progresses, it loses the ability to generate HFR learned from the source domain. In contrast, it can be observed that there is an impressive improvement in performance with the addition of the proposed modules.

Ablation study on NR Block. We perform ablation by keeping all our other modules intact while changing only the update methods. As shown in Tab. 5, batch normalization (BN) [18] and layer normalization (LN) [1] lead to a decrease in performance. The proposed NR Block update method significantly improves performance, with group normalization (GN) [57] being particularly effective.

The benefits of the adaptive threshold in RPS. We calculate the threshold for the reliable map between the student and teacher networks adaptively for each scene in the RPS module. Table 7 compares the adaptive thresholding method for the reliable map in the RPS module with a fixed-value thresholding approach.

The ablation study of mixing ratio λ in PMS. Table 6 presents an analysis of the mixing ratio between the samples in the memory bank and the generated videos in PMS.

Table 6. Performance based on minimum value of mixing ratio λ.

min(λ)	0.5	0.6	0.7	0.8	0.9
Cross	31.92	32.06	**32.13**	32.04	31.93
Continuous	33.30	33.49	**33.53**	33.42	33.36

Table 7. Analysis on the reliable threshold.

ρ	∞	0.2	0.1	0.05	0.01	Adaptive (Ours)
Cross	30.87	30.93	31.44	31.85	31.42	**32.13**
Continuous	32.21	32.64	33.02	32.97	32.78	**33.53**

Table 8. Inference time on different updating intervals. In the case of TTA-EVF, the inference time includes the training time as well.

S	T	Methods (TimeLens [51])	Time	PSNR	SSIM
B	H-close	No adapt	80 min	29.96	0.8413
		TTA-EVF (Interval=5)	81 min	31.44	0.8541
		TTA-EVF (Interval=3)	82 min	31.62	0.8571
		TTA-EVF (Interval=1)	88 min	**32.11**	**0.8598**

We select the mixing ratio λ uniformly, but it can be set with a minimum value. The best performance is observed when the minimum value, min(λ), is set to 0.7, demonstrating robustness to parameter variations.

6. Analysis of Updating Interval

We present the performance based on the updating interval in Table 8. Our TTA method is highly cost-effective despite performing training during the inference process. TTA-EVF, which updates only once during the multiple inferences when an video triplet is input, exhibits minimal time difference compared to simple inference. When the interval is increased, the time required is almost the same as simple inference, yet the quality of the generated videos is significantly higher. For example, with an interval of 5, there is only a 1 minute difference in total inference time, but the PSNR is improved by 1.48 dB.

7. Conclusion

In this paper, we present a Test-Time Adaptation (TTA) for event-based video frame interpolation. We effectively perform online TTA using the newly proposed modules in both cross-domain and continuous domain-shifting settings. Additionally, we acquire the ERDS dataset, containing high-quality RGB and event datasets with continuous distribution changes. Our TTA-EVF, despite online learning, shows minimal speed differences compared to simple inference and outperforms both supervised and unsupervised methods in cross-domain experiments, showcasing its effectiveness.

Acknowledgements. This research was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF2022R1A2B5B03002636) and the Challengeable Future Defense Technology Research and Development Program through the Agency For Defense Development (ADD) funded by the Defense Acquisition Program Administration (DAPA) in 2024 (No.912768601).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 8

[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. 1, 2

[4] Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-Jussa. Continual lifelong learning in natural language processing: A survey. *arXiv preprint arXiv:2012.09823*, 2020. 5

[5] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022. 2

[6] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 454–462, 2022. 2

[7] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022. 2

[8] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023. 2

[9] Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, and Kuk-Jin Yoon. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12492–12503, 2023. 2

[10] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19866–19877, 2023. 4

[11] Myungsub Choi, Janghoon Choi, Sungyong Baik, Tae Hyun Kim, and Kyoung Mu Lee. Scene-adaptive video frame interpolation via meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9444–9453, 2020. 2, 6, 7

[12] 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*, pages 10663–10671, 2020. 1

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2

[14] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022. 2

[15] 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. 1, 2

[16] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17804–17813, 2022. 1, 2, 4, 6, 7

[17] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022. 1

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 3, 8

[19] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021. 2

[20] 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 conference on computer vision and pattern recognition*, pages 9000–9008, 2018. 1, 2, 3, 6

[21] Shihao Jiang, Yao Lu, Hongdong Li, and Richard Hartley. Learning optical flow from a few matches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16592–16600, 2021.

[22] Rico Jonschkowski, Austin Stone, Jonathan T Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *European Conference on Computer Vision*, pages 557–572. Springer, 2020. 1, 2

[23] Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17745–17754, 2022. 2

[24] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18032–18042, 2023. 1, 3, 6, 7, 8

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neu-

ral networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 5

[27] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020. 2

[28] Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10615–10625, 2023. 2

[29] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 3

[30] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. *ECCV*, 2020. 2

[31] Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22952–22961, 2023. 2

[32] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820, 2021. 2

[33] Huaijia Lin Jiangbo Lu Liying Lu, Ruizheng Wu and Jiaya Jia. Video frame interpolation with transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[34] Weng Fei Low and Gim Hee Lee. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18335–18346, 2023. 2

[35] Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. Upflow: Upsampling pyramid for unsupervised optical flow learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1045–1054, 2021. 1, 2

[36] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14765–14775, 2022. 2

[37] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *European Conference on Computer Vision*, 2020. 1, 2

[38] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *International Conference on Computer Vision*, 2021. 1, 2

[39] Fitsum A Reda, Deqing Sun, Aysegul Dundar, Mohammad Shoeybi, Guilin Liu, Kevin J Shih, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Unsupervised video interpolation using cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 892–900, 2019. 2, 4

[40] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020. 2

[41] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S. Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4531–4540, 2021. 2

[42] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 2

[43] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[44] Chen Song, Qixing Huang, and Chandrajit Bajaj. E-cir: Event-enhanced continuous intensity recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7803–7812, 2022. 2

[45] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[46] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 6

[47] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 2

[48] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. *ArXiv*, abs/2203.10016, 2022. 2

[49] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 4

[50] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 1, 2

[51] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpola-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16155–16164, 2021. 1, 3, 6, 7, 8

[52] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. *arXiv preprint arXiv:2203.17191*, 2022. 1, 6

[53] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 2, 3, 6

[54] Yusheng Wang, Yunfan Lu, Ye Gao, Lin Wang, Zhihang Zhong, Yinqiang Zheng, and Atsushi Yamashita. Efficient video deblurring guided by motion magnitude. In *European Conference on Computer Vision*, pages 413–429. Springer, 2022. 4

[55] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based blurry frame interpolation under blind exposure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1588–1598, 2023. 2

[56] Song Wu, Kaichao You, Weihua He, Chen Yang, Yang Tian, Yaoyuan Wang, Ziyang Zhang, and Jianxing Liao. Video interpolation by event-driven anisotropic adjustment of optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[57] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4, 8

[58] Ruihao Xia, Chaoqiang Zhao, Meng Zheng, Ziyan Wu, Qiyu Sun, and Yang Tang. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21515–21524, 2023. 2, 4

[59] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 2

[60] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *NeurIPS*, 2019. 1, 2

[61] 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. 1

[62] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5

[63] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35: 38629–38642, 2022. 2

[64] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17765–17774, 2022. 2

[65] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022. 2

[66] Yizhe Zhang, Shubhankar Borse, Hong Cai, and Fatih Porikli. Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2339–2348, 2022. 2

[67] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3