

From Sharp to Blur: Unsupervised Domain Adaptation for 2D Human Pose Estimation Under Extreme Motion Blur Using Event Cameras

Youngho Kim*, Hoonhee Cho*, and Kuk-Jin Yoon
KAIST

{kmax2001, gnsngsgml, kjyoon}@kaist.ac.kr

Abstract

Human pose estimation is critical for applications such as rehabilitation, sports analytics, and AR/VR systems. However, rapid motion and low-light conditions often introduce motion blur, significantly degrading pose estimation due to the domain gap between sharp and blurred images. Most datasets assume stable conditions, making models trained on sharp images struggle in blurred environments. To address this, we introduce a novel domain adaptation approach that leverages event cameras, which capture high temporal resolution motion data and are inherently robust to motion blur. Using event-based augmentation, we generate motion-aware blurred images, effectively bridging the domain gap between sharp and blurred domains without requiring paired annotations. Additionally, we develop a student-teacher framework that iteratively refines pseudo-labels, leveraging mutual uncertainty masking to eliminate incorrect labels and enable more effective learning. Experimental results demonstrate that our approach outperforms conventional domain-adaptive human pose estimation methods, achieving robust pose estimation under motion blur without requiring annotations in the target domain. Our findings highlight the potential of event cameras as a scalable and effective solution for domain adaptation in real-world motion blur environments. Our project codes are available at <https://github.com/kmax2001/EvSharp2Blur>.

1. Introduction

Human pose estimation identifies key body joints or limbs, essential for applications in rehabilitation, sports analytics, and AR/VR systems. Despite the dynamic nature of human motion, most datasets assume stable conditions, providing only clean, well-conditioned data. As a result, models trained on such datasets struggle with motion blur due to the significant domain gap between sharp and blurred images. Conventional cameras rarely capture paired sharp-blurred

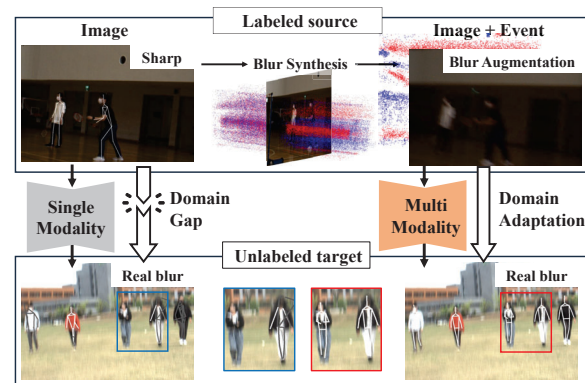


Figure 1. Image-based networks trained on sharp images face performance degradation in blurred images due to the domain gap. In contrast, our method uses event data for blur augmentation and achieves superior performance in the blurred domain through multi-modality adaptation.

images, making domain adaptation challenging. While dual-camera setups with beam splitters [16, 18] offer precise annotations, they require expertise and are impractical for general user use. Addressing motion blur without such complex systems remains a key challenge in pose estimation.

Unlike conventional cameras, event cameras [22, 77] operate at a high frame rate, making them resistant to motion blur. In contrast to standard cameras, which capture images at fixed intervals, event cameras detect per-pixel intensity changes, producing sparse but temporally rich data. Previous research has utilized event cameras for motion deblurring [11, 27, 63, 65, 78], image reconstruction [4, 68, 70, 79], and perceptions under low-light conditions [31, 39, 42–45, 73, 80]. However, efforts to bridge the motion blur domain gap, particularly in multi-person pose estimation, remain unexplored.

In this study, we focus on transferring knowledge from a sharp domain with ground-truth annotations to a motion-blurred environment where pose annotations are difficult to obtain. Unsupervised domain adaptation relies on generating reliable pseudo-labels [14, 73, 83] in the target domain. However, the large domain gap between sharp and blurred

*The first two authors contributed equally.

images makes this a challenging task, and traditional cameras alone lead to suboptimal adaptation. To address this, we introduce a novel method for adapting 2D multi-human pose estimation from a sharp domain to a blurred domain using event cameras, which capture continuous motion data over time. Unlike static sharp images that lack motion information, event cameras provide detailed motion data, enabling each time slice to capture the full pixel movement within that period. As illustrated in Fig. 1, we utilize this characteristic to generate motion-aware blurred images through event-based augmentation, using event cameras to bridge the gap between the two domains. By employing these synthesized motion-aware blurred images, we effectively reduce the domain discrepancy, allowing the model to maintain reasonable performance in the target domain even when trained exclusively on the source domain.

In addition, we focus on the complementary strengths of images and event data. Images capture dense spatial information, effectively preserving semantic details in sharp domains and often helping in blurred domains as well. In contrast, event cameras capture sparse motion cues along moving edges, enabling robust detection even under extreme motion but often suffering from false detections due to their lack of spatial density. To leverage the strengths of both modalities, we design a teacher network with a sub-network that takes different types of modalities as input and a refinement module that adaptively utilizes the strengths of each output of sub-networks. This module enables scene-specific pose estimation by balancing the spatial richness of images and the motion-invariant edge features of events, ultimately generating high-quality pseudo-labels for adaptation. Additionally, we propose a method that jointly employs a student-teacher framework for pseudo-label generation. We alternately adapt the student and teacher networks to the target domain, with the teacher network leveraging the student network to generate more reliable pseudo-labels for adaptation. In turn, the enhanced teacher network further trains the student network, creating a feedback loop that improves overall performance.

Our approach, the first to leverage event cameras to bridge the domain gap caused by motion blur, enables effective adaptation from the sharp domain to the blur domain without ground-truth annotations in the blur domain. During the adaptation process, our method minimizes performance degradation in the source domain, ultimately achieving superior results in both the sharp and blur domains.

2. Related Works

Domain-adaptive Human Pose Estimation. There are two paradigms in human pose estimation: bottom-up and top-down approaches. The top-down method detects the human first, and then estimates keypoint locations [7, 19, 51, 53, 61, 66, 69], while the bottom-up method detects keypoints first, and then groups them into poses [6, 9, 29, 30, 40, 52, 56, 62]

Training 2D human pose estimation models requires large labeled datasets, which are labor-intensive and time-consuming. To address this, synthetic datasets are often used [35, 46], but a domain gap remains, limiting generalization across datasets [23, 58]. To bridge this gap, data augmentation techniques are commonly applied to mitigate these gaps [1, 8, 17, 35, 46]. Despite advances in domain adaptation for human pose estimation, little has been explored in the challenging blurry domains. Estimation of pose in blurred images [82] is difficult due to the complexities of the augmentation needed to bridge the domain gap. Specifically, generating blur based on motion from a single image requires sophisticated methods that account for object motion, which is particularly challenging. To this end, we propose a novel event-based augmentation method that considers actual motion to address these challenges.

Pseudo-label Refinement. To train on an unlabeled dataset, teacher-student frameworks are often adopted, where teacher networks generate pseudo labels to guide student network training [2, 12, 15, 20, 28, 41, 72–74, 84]. To enhance the reliability of pseudo labels, various refinement techniques have been explored. For example, MixMatch [2] applies data augmentation at various levels and aggregates predictions to improve robustness. LEOD [72] employs a high confidential threshold to obtain more reliable pseudo labels and utilizes tracking-based post processing to filter out temporally inconsistent and noisy labels. SSPCM [28] used pseudo-label correction module that selects only consistent prediction from dual networks. We also propose a learning-based pseudo-label refinement module that leverages outputs from multiple models, each trained on a different modality.

Event-based Human Pose Estimation. Event cameras are robust to motion blur and have low latency, which has led to growing attempts to apply them to human pose estimation. Specifically, there have been efforts to estimate dynamic features such as hand pose [33, 34, 50, 60], leveraging the ability of event cameras to perform well in extreme lighting and blur conditions. Additionally, datasets and approaches for whole-body pose estimation [3, 48, 49, 75, 76, 85] are emerging. Recently, the first RGB and event dataset [16], including pose annotations for multi-human poses, has emerged, encompassing extreme blur and low-light conditions. Building on these recent efforts using event cameras to tackle the blur domain, we are the first to attempt bridging the gap between sharp and blur domains in multi-human pose estimation.

Event-based Domain Adaptation. Event cameras have varying distributions depending on the illumination and the device, and approaches for adapting models to these changing distributions [14, 26, 32, 36, 57] have been continually proposed. Additionally, there have been efforts [5, 10, 47, 64, 67, 83] to adapt from the image domain to the event domain to compensate for the lack of sensor data, as well as approaches to adapt perception models for

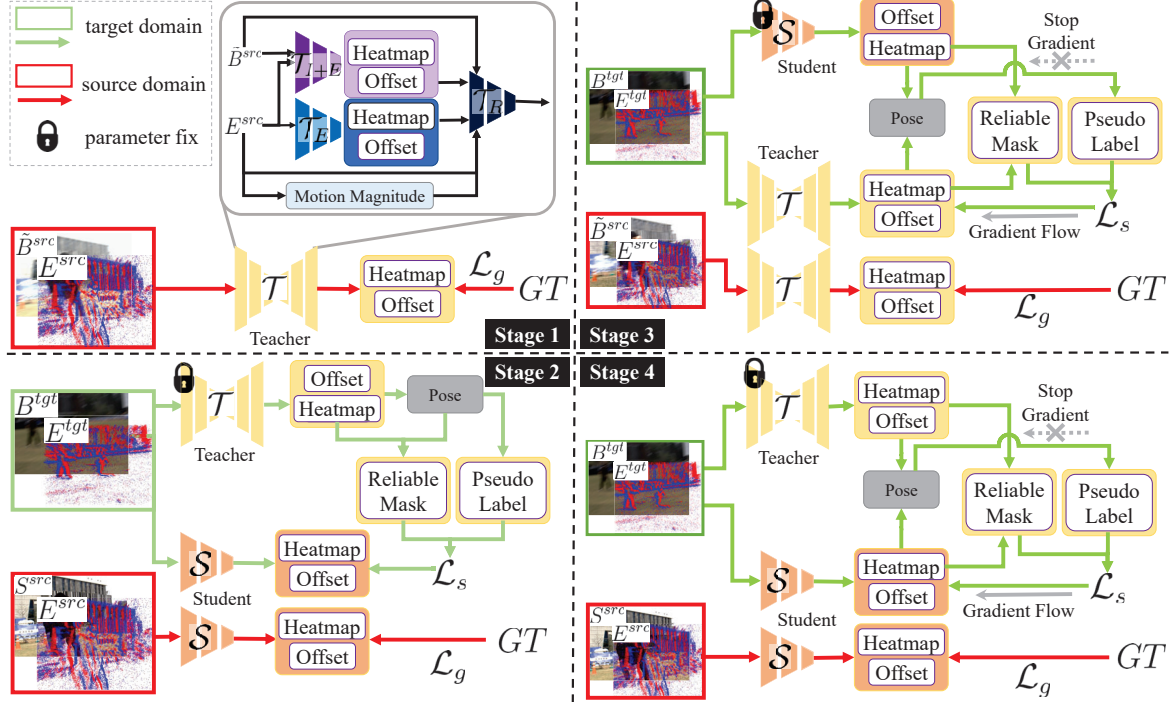


Figure 2. Overall framework of the proposed domain-adaptive human pose estimation. The proposed method consists of four sequential stages. Stage 1: Teacher Network Pre-training, Stage 2: Student Network Training with Pseudo-Labels, and Stages 3 and 4: Mutual Uncertainty Masking for the Teacher and Student Networks. GT denotes the ground-truth human pose.

low-light conditions [73]. As another new challenge, we use event cameras as a bridge between motion blur and sharp domains, and for the first time, we attempt to adapt to the blur domain using annotations only from the sharp domain.

3. Methods

3.1. Problem Definition and Overall Framework

Domain adaptive 2D human pose estimation aims to achieve high performance on a target domain with a different data distribution, using only labeled source domain data. We assume the source domain consists of sharp images, S^{src} , with annotations, while the target domain contains blur images, B^{tgt} , without annotations. Event data is present in both domains, with E^{src} representing events from the source domain and E^{tgt} from the target domain. Each image corresponds to a single frame, and our goal is to address multi-human 2D pose estimation.

Our network structure follows DEKR [24]. Each network simultaneously outputs a center heatmap \mathcal{H}_c , offset \mathcal{O} , and keypoint heatmaps \mathcal{H}_k , where k represents the index of each keypoint. The center heatmap detects the human’s center whose location is the local argmax of the center heatmap. Offsets, defined as $\mathcal{O} = \{p_c^i - p_1^i, p_c^i - p_2^i, \dots, p_c^i - p_k^i, \dots, p_c^i - p_K^i\}$, are displacement of each the human’s center p_c^i from keypoints location p_k^i where $p_k^i \in R^2$ represents the 2D location of the

k -th keypoints for the i -th person. The keypoint heatmap is used for scoring and ranking the regressed poses by averaging the heat value on each keypoint location. Following [24], supervised loss for heatmap and offset is defined as

$$\mathcal{L}_g = L_{\mathcal{H}} + \lambda_g \cdot L_{\mathcal{O}} \quad (1)$$

$L_{\mathcal{H}}$ means a MSE loss of the predicted heatmap and $L_{\mathcal{O}}$ means a smooth L1 loss of the offset map. λ_g is a hyper-parameter and set as 0.03 in our experiment.

As shown in Fig. 2, our overall training process is divided into four main stages. **(1) Teacher Network Pretraining:** We train a high-capacity teacher network that leverages multi-modal learning. To ensure superior performance on the target domain using only source domain data, we propose event-based augmentation techniques. **(2) Student Network Training with Pseudo-Labels:** Using the pretrained teacher network, we generate pseudo-labels to train the student network \mathcal{S} . The goal is to enable \mathcal{S} to learn effectively from both the source and target domains. **(3) Mutual Uncertainty Masking for Teacher Network:** The trained student and teacher networks are then used together to mutually mask uncertain predictions, generating more reliable pseudo-labels. **(4) Mutual Uncertainty Masking for Student Network:** The enhanced teacher is used again to generate pseudo-labels with student network, further refining the student network’s performance.

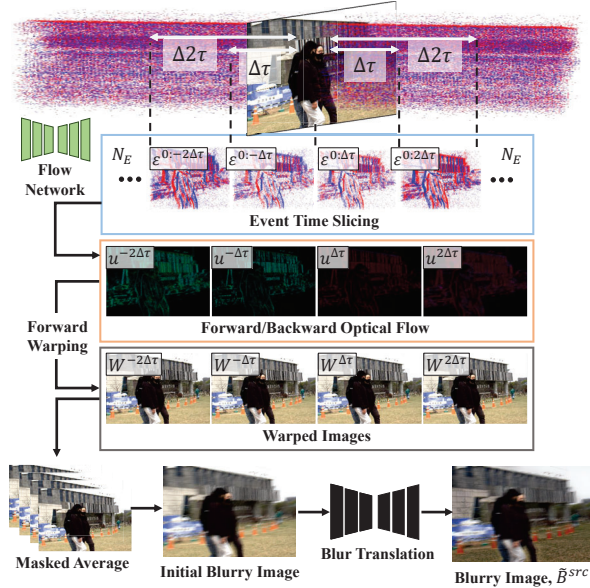


Figure 3. Motion-aware event-based blur augmentation.

3.2. Teacher Network Pretraining

In the target domain, where labels are unavailable, recent studies [1, 8] have used teacher networks to generate pseudo-labels for effective training on unlabeled data. A higher-capacity of teacher network and robustness to domain shifts generate better pseudo-labels. We propose a robust teacher network trained with motion-aware blur augmentation using event data to ensure strong performance in the target domain, even when trained solely on the source domain. Additionally, the teacher network incorporates multiple sub-networks that use various modalities, with a refinement network.

Motion-aware Event-based Blur Augmentation. Our setup deals with the scenario of a single image, similar to the domain-adaptive multi-human pose estimation [1, 35]. Generating blur from a single image is extremely challenging. As an alternative, one can randomly generate a blur kernel and apply it to the image to simulate blur. However, this approach does not follow the actual principles of motion blur. Since it is not generated based on any motion, it is particularly ineffective in representing blur caused by object movement. To solve this problem, we propose the blur augmentation method that considers the actual motion by leveraging event data. Real-world motion blur images [18, 37, 38, 59] are formed by the continuous accumulation of pixel information during the exposure time. To replicate this, we compute pixel movement from events and shift pixels accordingly. For this, we train an event-based optical flow network with a self-supervised loss [13, 21, 54]. As shown in Fig. 3, we estimate motion by slicing the event data during a time duration longer than the exposure time of the sharp image, with the center timestamp of the sharp image serving as a reference.

We divide the event slice set into $2N_E$ seg-

ments over $\Delta\tau$ and estimate the flow for each slice: $\{u^{(t\Delta\tau)} \mid t = -N_E, \dots, N_E\}$, based on the corresponding event sequences: $\{\varepsilon^{(0:t\Delta\tau)} \mid t = -N_E, \dots, N_E\}$. The estimated flows are then used for forward warping on a sharp reference image to generate the warped images: $\{W^{(t\Delta\tau)} \mid t = -N_E, \dots, N_E\}$. Forward warping often introduces holes, so instead of averaging across all images, we handle these gaps as follows:

$$\tilde{B}^{src}(x, y) = \frac{\sum_t W_t(x, y) \cdot \delta_t(x, y)}{\sum_t \delta_t(x, y) + \epsilon} \quad (2)$$

where $W_t(x, y)$ denotes the intensity of the t -th warped image at location (x, y) . If the t -th sharp image contains a hole at (x, y) , then $\delta_t(x, y) = 0$; otherwise, $\delta_t(x, y) = 1$. The small constant ϵ prevents division by zero.

This process still generates a discrete blur image. To make it more realistic, we apply a blur translation loss [55] to the generated blur in order to reduce the distribution gap between discrete and continuous blurry images.

Multi-modal teacher network. Event and image data have distinct characteristics. Event-based networks are robust to motion blur, maintaining performance across domains. In contrast, image-based networks are highly sensitive to motion blur but capture rich intensity information. When trained with annotated blur images and combined with event data, network using both image and event outperforms purely image- and event-based approaches.

To leverage the advantages of each modality, we design the overall teacher network, \mathcal{T}_M , as a structure composed of sub-teacher networks, each based on different modalities. They are constructed based on [16, 24], where $\mathcal{M} \in \{I + E, E\}$ means the modality of the input of the network. As shown in Stage 1 of Fig. 2, our teacher network consists of two sub-networks: an only event-based network, \mathcal{T}_E , and an image-event fusion network, \mathcal{T}_{I+E} . To combine the results of these two networks, we use an additional refinement network, \mathcal{T}_R . The refinement network generates the final heatmap and offset using image and event inputs, along with those from each sub-network. Since the teacher network is used only during training, we enhance results by providing additional guidance. Specifically, the refinement network receives the magnitude of optical flow, computed from events. The refinement network, \mathcal{T}_R , concatenates all inputs along the channel dimension and fuses them through a channel attention mechanism [63, 81].

To ensure effective adaptation of the teacher network to the blur domain, Stage 1 excludes sharp images. Instead, the teacher network is trained with synthesized blur images generated through augmentation, using source domain annotations and supervised by \mathcal{L}_g (Eq. (1)) during this process.

3.3. Student Network Training with Pseudo-Labels

In Stage 2, the goal is to train the student network using the accurate annotations from the source domain and pseudo-labels from teacher network in the target domain. The student network takes image and event data as inputs and has the same architecture as \mathcal{T}_{I+E} . Using the teacher network trained in Stage 1, which is partially adapted to the blur domain through augmentation and takes advantages of multi-modal data through refinement, we generate pseudo-labels for the target dataset.

Firstly, we estimate center location from center heatmap, \mathcal{H}_c^T and generate each pose $P_i^T = \{p_k^i, p_c^i\}_{k \in \{1, \dots, K\}}$ using offset, \mathcal{O}_k^T of each keypoint. p_k^i means keypoint location of k -th keypoint of i -th pose. Poses, $\{P_i^T\}$, which is set of each pose are then filtered out based on non-maximum suppression (NMS) [25]. However, the performance can degrade significantly when noisy or incorrect labels are provided during training. To avoid this, we introduce pixel-wise masking, M , to calculate the loss function. By averaging keypoint heatmap values on each keypoint, we calculate confidence score of each pseudo pose as follows:

$$C(P_i) = \mathcal{H}_c^T(p_c^i) * \sum_{k=1}^K \mathcal{H}_k^T(p_k^i) / K \quad (3)$$

Masking is applied when the confidence score is below th , and the confidence-based heatmap mask $M^{\mathcal{H}}$ is defined as:

$$d_k(x, y)^{\mathcal{H}} := \{i | (x, y) \in \text{near}(p_k^i)\} \quad (4)$$

$$M_k(x, y)^{\mathcal{H}} = \prod_{i \in d_k(x, y)^{\mathcal{H}}} m, \quad m = \begin{cases} 1, & C(P_i) \geq th \\ 0, & \text{else} \end{cases} \quad (5)$$

If $d_k(x, y)^{\mathcal{H}} = \emptyset$, the pixel (x, y) is considered as a background and $M_k(x, y)^{\mathcal{H}}$ is set to 0.1, following previous work [24]. The mask $M_k(x, y)^{\mathcal{H}}$ filters out where unreliable pseudo pose lies not to calculate loss of the keypoint k at (x, y) . This helps to exclude unreliable pixels from both the heatmap and offset, ensuring stable training in the target domain. $M_k(x, y)^{\mathcal{O}}$ is defined below.

$$M_k(x, y)^{\mathcal{O}} = \frac{1}{\min(Z_i)} \quad (6)$$

$Z_i = \{\sqrt{H_i^2 + W_i^2} | (x, y) \in \text{near}(p_c^i)\}$ is the set of the size of the corresponding person instance, and H_i and W_i are the height and width of the instance box [24]. If $Z_i = \emptyset$, $M_k(x, y)^{\mathcal{O}} = 0$. To compute the loss during student network training with the mask applied, where \mathcal{O} , \mathcal{H} are offset prediction and heatmap prediction of student network, we use the following approach:

$$L_{\mathcal{H}} = M^{\mathcal{H}} * \|\mathcal{H} - \mathcal{H}^T\|_2^2 \quad (7)$$

$$L_{\mathcal{O}} = M^{\mathcal{O}} * \text{Smooth}_{L_1}(\mathcal{O} - \mathcal{O}_k^T) \quad (8)$$

3.4. Mutual Uncertainty Masking

Teacher Network Training. Through Stages 1 and 2, we obtain a student network partially adapted to the target domain by using pseudo-labels generated by teacher network. These pseudo-labels depend on the performance of the teacher network so using a higher-performing teacher network can generate higher-quality pseudo-labels. However, since the pre-trained teacher network has not seen real blur yet in the target domain, there is still room for performance improvement through adaptation to the target domain.

In the blur domain, noise often causes false detections and missed detections, making it unreliable to fully trust a single network. Therefore, we propose mutual uncertainty masking, where the confidence scores from both the teacher and student networks are considered to mask out unreliable regions. Specifically, we first generate teacher poses, $\{P_i^T\}$, from the teacher network and student poses, $\{P_i^S\}$, from the student network. In Stages 3 and 4, we compute another confidence score, $C'(P_i)$, for the poses of each network based on the heatmap of the other network. The equation below calculates the confidence scores using the student network, \mathcal{S} :

$$C'(P_i^S) = \mathcal{H}_c^T(p_c^i) * \sum_{k=1}^K \mathcal{H}_k^T(p_k^i) / K \quad (9)$$

$$C(P_i^S) = \mathcal{H}_c^S(p_c^i) * \sum_{k=1}^K \mathcal{H}_k^S(p_k^i) / K \quad (10)$$

For scores of poses by \mathcal{T} , $C(P_i^T)$ should be calculated with \mathcal{H}_c^T , \mathcal{H}_k^T and $C'(P^T)$ should be calculated with \mathcal{H}_c^S , \mathcal{H}_k^S .

These two sets of poses are combined and passed through non-maximum suppression (NMS) to generate the final poses. For these final poses, both heatmap mask and offset mask are calculated based on the confidence from both networks. If either C' or C is lower than mutual masking threshold, th' , we mask out that region.

$$M'_k(x, y)^{\mathcal{H}} = \prod_{i \in d_k(x, y)^{\mathcal{T}}} m^{\mathcal{T}} \times \prod_{j \in d_k(x, y)^{\mathcal{S}}} m^{\mathcal{S}} \quad (11)$$

$$m^{\mathcal{T}} = \begin{cases} 1, & \min(C(P_i^T), C'(P_i^T)) \geq th' \\ 0, & \text{else} \end{cases}$$

$$m^{\mathcal{S}} = \begin{cases} 1, & \min(C(P_i^S), C'(P_i^S)) \geq th' \\ 0, & \text{else} \end{cases}$$

We can calculate $M'_k(x, y)^{\mathcal{O}}$ in the same way as $M'_k(x, y)^{\mathcal{H}}$, as done in the previous section (Sec. 3.3). The mutual mask M' is applied to the final poses to exclude unreliable predictions when calculating the loss, \mathcal{L}_s between pseudo label and prediction, similar to (Eqs. (1), (7), and (8)) in the previous stage. This masking strategy enhances the

Table 1. Multi-person pose estimation evaluated on the EHPT-XC dataset. To showcase the robustness of method beyond its performance in the target domain, we also report results in the source domain. The methods in top six rows represent oracle networks that leverage annotations from the target domain. ‘-’ indicates that evaluation was not performed, as the teacher network did not utilize sharp images of the source domain for training. Bold and underline indicate the best and second-best performances among networks trained only on the source domain, respectively.

Domain	Training Labels		Source (Sharp)		Target (Blur)		Average	
	Sharp	Blur	mAP@0.5:0.95	mAR@0.5:0.95	mAP@0.5:0.95	mAR@0.5:0.95	mAP@0.5:0.95	mAR@0.5:0.95
Base (I) [24]		✓	48.2	55.8	36.1	47.1	42.2	51.4
Base (E) [24]		✓	34.5	45.6	34.5	45.6	34.5	45.6
Base (I+E) [16]		✓	55.1	60.6	54.2	59.8	54.6	60.2
Base (I) [24]	✓	✓	68.9	73.4	53.1	61.4	61.0	67.4
Base (E) [24]	✓	✓	45.8	56.1	45.8	56.1	45.8	56.1
Base (I+E) [16]	✓	✓	74.5	77.9	58.8	65.1	66.6	71.5
Base (I) [24]	✓		62.5	67.2	28.6	32.7	45.6	50.0
Base (E) [24]	✓		43.5	49.7	43.5	49.7	43.5	49.7
Base (I+E) [16]	✓		67.9	72.0	37.5	40.5	<u>52.7</u>	<u>56.2</u>
DualTeacher (I) [1]	✓		45.9	52.6	24.4	29.8	35.2	41.2
DualTeacher (E) [1]	✓		33.5	41.6	33.5	41.6	33.5	41.6
DualTeacher (I+E) [1]	✓		60.1	64.7	34.6	37.5	47.4	51.1
UDA-HE (I) [35]	✓		46.6	53.4	18.0	23.3	32.3	38.4
UDA-HE (E) [35]	✓		29.8	38.5	29.8	38.5	29.8	38.5
UDA-HE (I+E) [35]	✓		61.8	66.6	36.4	40.6	49.1	53.6
Ours-Teacher	✓		-	-	52.9	58.2	-	-
Ours	✓		<u>64.2</u>	<u>68.1</u>	<u>51.6</u>	<u>57.5</u>	57.9	62.8

performance of the teacher network during adaptation to the target domain, improving its ability to handle real blur in the target domain.

Student Network Training. We now proceed to retrain the student network using the teacher network that has been improved and well-adapted to the target domain. As before, we use the mutual uncertainty masking (Eqs. (9) and (11)), but this time, instead of training the teacher network with pseudo label, we focus on training the student network. This ensures that the student network benefits from the improved adaptation of the teacher network to the target domain while considering the uncertainties introduced by both networks.

4. Experiments

4.1. Dataset

EHPT-XC Dataset. The EHPT-XC [16] dataset consists of 158 diverse sequences, each containing pixel-wise aligned and temporally synchronized event streams, sharp images and blurry images captured using the triplet camera system. The dataset includes 38K 2D keypoints, with 14 keypoints per person, along with bounding boxes and track IDs. For the unsupervised domain adaptation setting, we conceptually divide EHPT-XC into three subsets: train-source, train-target, and test. The train-source subset consists of 47 sequences, each containing 100 sharp image-event stream pairs with annotations. The train-target subset contains the images and events data, but with blurry, unlabeled images. The test subset comprises 26 sequences, each providing synchronized sharp images, blurry images, and event streams.

4.2. Implementation Details

The input image size is set to 512×512 . EHPT-XC dataset training is conducted on 2 TITAN RTX GPUs and batch size is 5. The training times for Stages 1–4 are 8 hours, 12 hours, 12 hours, and 12 hours, respectively. The learning rate is $1e-3$ during total 100 epochs. For blur augmentation, we used $N_E=5$. In addition to motion-aware augmentation, we applied both sharp and blurry images with random rotation($-30 \sim 30$), horizontal flip with 0.5 probability, random scale ($0.75 \sim 1.5$), random translation($-40 \sim 40$ pixels) in both x, y directions. During pose estimation, we select top 30 center point locations from center heatmap and filter out center of which heat value is smaller than the keypoint threshold, 0.03 in this experiments. We set keypoints per human, $K = 14$. $near(p)$ is set to correspond to a square region with a side length of 8 around the 2D location p .

4.3. Experimental Results

We establish a baseline by using the same network architecture as our model while varying only the input modality. For some baseline settings, training was conducted using both sharp images and target-domain labels, which we refer to as the oracle setting. Additionally, to compare our approach with existing domain-adaptive human pose estimation methods, we evaluated DualTeacher [1] and UDA-HE [35]. Since these are image-based methods, we extend the experiments by incorporating various modalities. Since DualTeacher [1]’s original augmentation is tailored for low-light conditions, we substituted it with standard augmentations.

Quantitative Results. Table 1 presents a quantitative comparison between the baselines, other methods, and our ap-

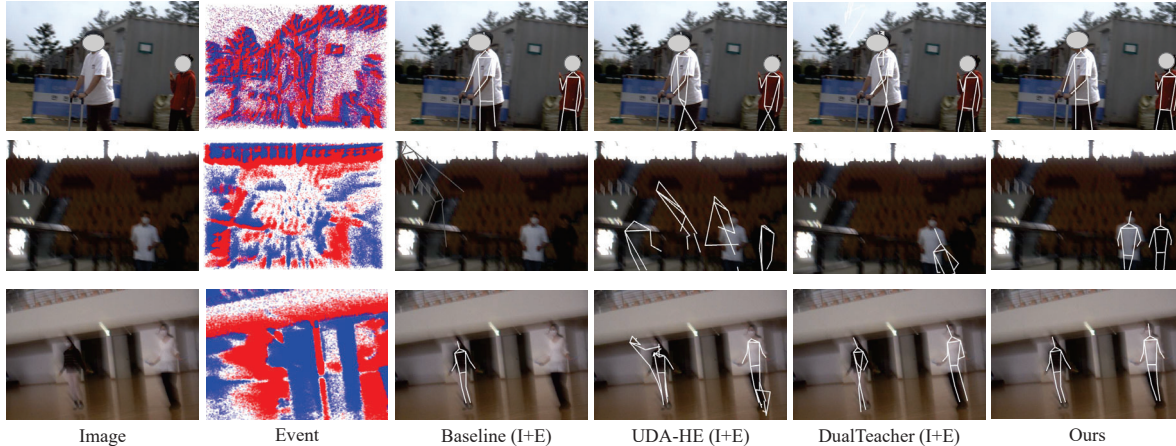


Figure 4. Comparison of UDA-HE [35], DualTeacher [1], baselines, and our method. The first row represents a scene with minimal blur, while the second and third rows depict scenes with severe blur.

Table 2. Ablation study of the proposed method. MBA denotes the motion-aware event-based blur augmentation. Since Stage 3 trains only the teacher network, we report the results of performing both Stage 3 and Stage 4 together.

	MBA	Stage 2	Stage 3 & 4	Target (Blur)	
				mAP	mAR
(A)				37.5	40.5
(B)	✓			40.4	46.5
(C)		✓		49.1	54.3
(D)	✓	✓		50.4	55.5
(E)	✓	✓	✓	51.6	57.5

proach. Among the baselines, image-based estimation suffers the largest accuracy drop, while event-based estimation is more resilient. The event modality has the smallest domain gap due to its high-resolution motion information, but it underperforms in the sharp domain. Domain-adaptive methods [1, 35] show improved performance over baselines in both multi-modal and image-based settings. However, since these methods were not originally designed for the blurred domain, the significant data distribution gap leads to sub-optimal performance. In contrast, our approach effectively bridges the domain gap early on through augmentation and a multi-modal teacher network, leveraging pseudo-label refinement to achieve superior results. Notably, our method achieves performance comparable to the multi-modal oracle. **Qualitative Results.** Figure 4 compares our method against multi-modal baselines and domain-adaptive methods. Baselines without domain adaptation experience significant performance degradation as blur intensifies. While domain-adaptive methods achieve some level of pose estimation, they produce a high number of false positives. In contrast, our approach demonstrates superior performance regardless of the blur intensity, as qualitatively observed.

5. Ablation Study and Analysis

Effectiveness of the modules. Table 2 shows the ablation study results for different components, evaluating adaptation

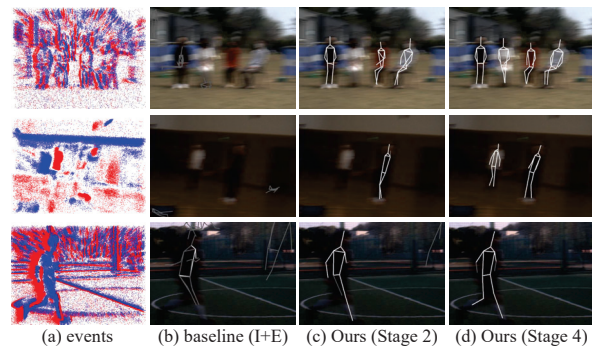


Figure 5. Qualitative comparison involving performance of each stage of our methods with baseline (image + event), stage 2 and stage 4 model. Ours refers to the student network.

Table 3. Effectiveness of motion-aware event-based blur augmentation. We report the result of the teacher network from Stage 1. Note that teacher network is trained on only source domain.

Method	Target (Blur)	
	mAP	mAR
Baseline	46.5	50.4
+ Motion Aug.	46.7	51.9
+ Motion Aug. + Blur Translation	46.9	52.2

performance in the target domain as each proposed module is progressively added. The results confirm that each module and training strategy contributes to overall performance improvement. Notably, combining augmentation with pseudo-label-based learning significantly enhances accuracy, while additional refinement in Stages 3 and 4 further improves final performance. The improvement in target domain adaptation through pseudo-label refinement across stages is more clearly illustrated in Fig. 5.

Effectiveness of the motion-aware augmentations As shown in Table 3, compared with baseline, motion augmentation improves mAP by 0.4, and when combined with blur translation, it achieves a total improvement of 1.8.

Table 4. Analysis of the teacher network components. The results of the teacher network \mathcal{T} in Stage 1 are presented.

	Modality	Target (Blur)	
		mAP	mAR
w/o. \mathcal{T}_R	\mathcal{T}_I	30.5	37.2
	\mathcal{T}_E	43.5	49.7
	\mathcal{T}_{I+E}	40.4	46.5
w. \mathcal{T}_R	\mathcal{T}_I	32.7	37.0
	\mathcal{T}_E	43.1	48.8
	\mathcal{T}_{I+E}	39.6	46.2
	$\mathcal{T}_{I+E} + \mathcal{T}_E$ (Ours)	46.9	52.2

Table 5. Analysis of the method for merging the outputs of the sub-teacher networks in the teacher network. All methods use \mathcal{T}_{I+E} and \mathcal{T}_E as sub-teachers. The results of the teacher network \mathcal{T} in Stage 1 are presented.

Method	Target (Blur)	
	mAP	mAR
Mean Heat Map	45.4	50.8
Spatial Attention [71]	41.9	48.3
Refinement Network	46.9	52.2

Table 6. Hyper-parameter analysis of the masking threshold. th and th' denote the thresholds for Stage 2 and Stage 4, respectively.

	Stage 2		Stage 4		
	th	mAP	mAR	th'	mAP
0.05	47.2	52.8	0.05	49.7	56.3
0.1	50.4	55.5	0.1	51.6	57.5
0.2	47.7	52.5	0.2	49.7	56.9
0.3	45.7	51.0	0.3	50.3	56.9

Analysis of teacher network components. The experimental results on the components of the teacher network are presented in Table 4. The results indicate that relying on a single sub-teacher component fails to fully exploit the refinement module, as each modality exhibits strengths and weaknesses depending on the scene conditions. Our teacher network consists of a multi-modal (I+E) sub-teacher, an event (E) sub-teacher, and a refinement module, allowing it to leverage the benefits of multi-modal learning while effectively handling noisy labels. As a result, it achieves the highest performance.

Analysis of the refinement network. To refine the final output based on the results of the sub-teacher networks, we design and incorporate a refinement network. As alternatives, one could average the heatmaps from both networks or apply spatial attention [71] mechanisms. However, as shown in Table 5, our method, which computes high-representation attention across channels for fusion, achieves the best performance among all approaches.

Threshold Analysis. Table 6 presents the performance variations based on the threshold for training the student network in Stages 2 and 4. Notably, when relying on a single-teacher network, the model becomes more sensitive to threshold selection, leading to performance fluctuations. In contrast, with a dual-network framework, the filtering process produces more reliable pseudo-labels, resulting in greater robustness

Table 7. Performance of image-only student trained with an image+event teacher network at Stage 4.

Domain	Training Labels		Source (Sharp)		Target (Blur)		Average	
	Sharp	Blur	mAP	mAR	mAP	mAR	mAP	mAR
Base (I)		✓	48.2	55.8	36.1	47.1	42.2	51.4
Base (I)	✓		62.5	67.2	28.6	32.7	45.6	50.0
Ours (I)	✓		63.2	68.1	46.0	52.7	54.6	60.4

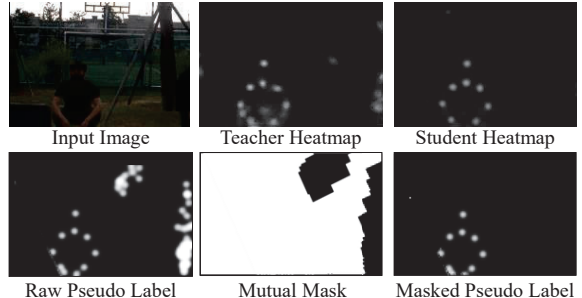


Figure 6. Pseudo label with mutual uncertainty masking.

across a wider range of thresholds.

Effectiveness of mutual uncertainty masking. Figure 6 illustrates the effectiveness of mutual uncertainty masking. In many cases, either the teacher or the student network predicts the correct answer while the other produces an incorrect one. Pseudo-labeling based on a single network is vulnerable to such inconsistencies. However, our approach can extract reliable masking even from the student heatmap, effectively filtering out uncertain regions in the teacher heatmap, which has a higher capacity. As a result, the masked pseudo-labels become more accurate, eliminating uncertain areas and improving overall reliability.

Analysis of using only image data at test time. To apply pose estimation at real-world, aligned image-event data can be hard to acquired. So we investigated the performance using only image data at test time. We trained a final student network using only image data with the pseudo labels acquired from the multi-modal teacher network. As shown in Table 7, without event data at test time, our framework shows that the proposed method can achieve better performance from the image-only baseline.

6. Conclusion

In this paper, we introduce event cameras for the first time to facilitate domain-adaptive human pose estimation from a sharp domain to a motion-blurred domain. Without relying on target-domain annotations, we propose an effective adaptation strategy leveraging motion-aware event-based blur augmentation and dual network-based mutual uncertainty masking. Our approach successfully bridges the domain gap, achieving performance comparable to that of an oracle model trained with target-domain annotations. Our work highlights the potential of event cameras for robust human pose estimation and their applicability in dynamic environments.

7. Acknowledgments.

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636), and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, AI Research Hub Project).

References

- [1] Y. Ai, Y. Qi, B. Wang, Y. Cheng, X. Wang, and R. T. Tan. Domain-adaptive 2d human pose estimation via dual teachers in extremely low-light conditions. In *European Conference on Computer Vision*, pages 221–239. Springer, 2024. [2](#), [4](#), [6](#), [7](#)
- [2] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. [2](#)
- [3] E. Calabrese, G. Taverni, C. Awai Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, and T. Delbruck. Dhp19: Dynamic vision sensor 3d human pose dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. [2](#)
- [4] M. Cannici and D. Scaramuzza. Mitigating motion blur in neural radiance fields with events and frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9286–9296, 2024. [1](#)
- [5] M. Cannici, C. Plizzari, M. Planamente, M. Ciccone, A. Bottino, B. Caputo, and M. Matteucci. N-rod: A neuromorphic dataset for synthetic-to-real domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1342–1347, 2021. [2](#)
- [6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. [2](#)
- [7] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. [2](#)
- [8] Y. Chen, L. Zhao, Y. Xu, H. Zu, X. An, and G. Li. Domain adaptive pose estimation via multi-level alignment. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. [2](#), [4](#)
- [9] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. [2](#)
- [10] H. Cho, J. Cho, and K.-J. 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](#)
- [11] H. Cho, Y. Jeong, T. Kim, and K.-J. 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. [1](#)
- [12] H. Cho, H. Kim, Y. Chae, and K.-J. 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. [2](#)
- [13] H. Cho, J.-Y. Kang, and K.-J. Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *European Conference on Computer Vision*, pages 294–314. Springer, 2024. [4](#)
- [14] H. Cho, T. Kim, Y. Jeong, and K.-J. Yoon. Tta-evf: test-time adaptation for event-based video frame interpolation via reliable pixel and sample estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25701–25711, 2024. [1](#), [2](#)
- [15] H. Cho, S.-H. Yoon, H. Kweon, and K.-J. Yoon. Finding meaning in points: Weakly supervised semantic segmentation for event cameras. In *European Conference on Computer Vision*, pages 266–286. Springer, 2024. [2](#)
- [16] H. Cho, T. Kim, Y. Jeong, and K.-J. Yoon. A benchmark dataset for event-guided human pose estimation and tracking in extreme conditions. *Advances in Neural Information Processing Systems*, 37:134826–134840, 2025. [1](#), [2](#), [4](#), [6](#)
- [17] V. Crescitelli, A. Kosuge, and T. Oshima. Poison: Human pose estimation in insufficient lighting conditions using sensor fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–8, 2020. [2](#)
- [18] P. Duan, B. Li, Y. Yang, H. Lou, M. Teng, Y. Ma, and B. Shi. Eventaid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *arXiv preprint arXiv:2312.08220*, 2023. [1](#), [4](#)
- [19] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017. [2](#)
- [20] H.-S. Fang, G. Lu, X. Fang, J. Xie, Y.-W. Tai, and C. Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310*, 2018. [2](#)
- [21] G. Gallego, H. Rebecq, and D. Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018. [4](#)
- [22] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. [1](#)
- [23] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. [2](#)
- [24] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14676–14686, 2021. [3](#), [4](#), [5](#), [6](#)

- [25] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4507–4515, 2017. 5
- [26] Y. Hu, T. Delbruck, and S.-C. Liu. Learning to exploit multiple vision modalities by using grafted networks. In *European Conference on Computer Vision*, pages 85–101. Springer, 2020. 2
- [27] K. Huang, S. Zhang, J. Zhang, and D. Tao. Event-based simultaneous localization and mapping: A comprehensive survey. *ArXiv*, abs/2304.09793, 2023. URL <https://api.semanticscholar.org/CorpusID:258213006>. 1
- [28] L. Huang, Y. Li, H. Tian, Y. Yang, X. Li, W. Deng, and J. Ye. Semi-supervised 2d human pose estimation driven by position inconsistency pseudo label correction module. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 693–703, 2023. 2
- [29] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 34–50. Springer, 2016. 2
- [30] U. Iqbal and J. Gall. Multi-person pose estimation with local joint-to-person associations. In *Computer Vision—ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II 14*, pages 627–642. Springer, 2016. 2
- [31] Y. Jeong, H. Cho, and K.-J. Yoon. Towards robust event-based networks for nighttime via unpaired day-to-night event translation. In *European Conference on Computer Vision*, pages 286–306. Springer, 2024. 1
- [32] D. Jian and M. Rostami. Unsupervised domain adaptation for training event-based networks using contrastive learning and uncorrelated conditioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18721–18731, 2023. 2
- [33] J. Jiang, J. Li, B. Zhang, X. Deng, and B. Shi. Evhandpose: Event-based 3d hand pose estimation with sparse supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [34] J. Jiang, X. Zhou, B. Wang, X. Deng, C. Xu, and B. Shi. Complementing event streams and rgb frames for hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24944–24954, 2024. 2
- [35] D. Kim, K. Wang, K. Saenko, M. Betke, and S. Sclaroff. A unified framework for domain adaptive pose estimation. In *European Conference on Computer Vision*, pages 603–620. Springer, 2022. 2, 4, 6, 7
- [36] J. Kim, I. Hwang, and Y. M. 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
- [37] T. Kim, H. Cho, and K.-J. Yoon. Cross-modal temporal alignment for event-guided video deblurring. *arXiv preprint arXiv:2408.14930*, 2024. 4
- [38] T. Kim, H. Cho, and K.-J. Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24966–24976, 2024. 4
- [39] T. Kim, J. Jeong, H. Cho, Y. Jeong, and K.-J. Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, pages 433–451. Springer, 2024. 1
- [40] S. Kreiss, L. Bertoni, and A. Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019. 2
- [41] D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013. 2
- [42] H. Li, J. Wang, J. Yuan, Y. Li, W. Weng, Y. Peng, Y. Zhang, Z. Xiong, and X. Sun. Event-assisted low-light video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3250–3259, 2024. 1
- [43] G. Liang, K. Chen, H. Li, Y. Lu, and L. Wang. Towards robust event-guided low-light image enhancement: a large-scale real-world event-image dataset and novel approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23–33, 2024.
- [44] J. Liang, Y. Yang, B. Li, P. Duan, Y. Xu, and B. Shi. Coherent event guided low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10615–10625, 2023.
- [45] H. Liu, S. Peng, L. Zhu, Y. Chang, H. Zhou, and L. Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 1
- [46] A. Martínez-González, M. Villamizar, O. Canévet, and J.-M. Odobez. Investigating depth domain adaptation for efficient human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2
- [47] N. Messikommer, D. Gehrig, M. Gehrig, and D. Scaramuzza. Bridging the gap between events and frames through unsupervised domain adaptation. *IEEE Robotics and Automation Letters*, 7(2):3515–3522, 2022. 2
- [48] C. Millerdurai, H. Akada, J. Wang, D. Luvizon, C. Theobalt, and V. Golyanik. Eventego3d: 3d human motion capture from egocentric event streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1186–1195, 2024. 2
- [49] C. Millerdurai, H. Akada, J. Wang, D. Luvizon, A. Pagani, D. Stricker, C. Theobalt, and V. Golyanik. Eventego3d++: 3d human motion capture from a head-mounted event camera. *arXiv preprint arXiv:2502.07869*, 2025. 2
- [50] J. Nehvi, V. Golyanik, F. Mueller, H.-P. Seidel, M. Elgharib, and C. Theobalt. Differentiable event stream simulator for non-rigid 3d tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1302–1311, 2021. 2

- [51] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 2
- [52] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 2
- [53] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4903–4911, 2017. 2
- [54] F. Paredes-Vallés and G. C. de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. *arXiv preprint arXiv:2009.08283*, 2020. 4
- [55] B.-D. Pham, P. Tran, A. Tran, C. Pham, R. Nguyen, and M. Hoai. Blur2blur: Blur conversion for unsupervised image deblurring on unknown domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2804–2813, 2024. 4
- [56] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 2
- [57] M. Planamente, C. Plizzari, M. Cannici, M. Ciccone, F. Strada, A. Bottino, M. Matteucci, and B. Caputo. Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *IEEE Robotics and Automation Letters*, 6(4):6616–6623, 2021. 2
- [58] M. Ragab, E. Eldele, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li. Self-supervised autoregressive domain adaptation for time series data. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):1341–1351, 2022. 2
- [59] J. Rim, H. Lee, J. Won, and S. Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 4
- [60] V. Rudnev, V. Golyanik, J. Wang, H.-P. Seidel, F. Mueller, M. Elgharib, and C. Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12385–12395, 2021. 2
- [61] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2
- [62] K. Sun, Z. Geng, D. Meng, B. Xiao, D. Liu, Z. Zhang, and J. Wang. Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates. *arXiv preprint arXiv:2006.15480*, 2020. 2
- [63] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European conference on computer vision*, pages 412–428. Springer, 2022. 1, 4
- [64] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 2
- [65] Z. Sun, X. Fu, L. Huang, A. Liu, and Z.-J. Zha. Motion aware event representation-driven image deblurring. In *European Conference on Computer Vision*, pages 418–435. Springer, 2024. 1
- [66] Z. Tian, H. Chen, and C. Shen. Directpose: Direct end-to-end multi-person pose estimation. *arXiv preprint arXiv:1911.07451*, 2019. 2
- [67] L. Wang, Y. Chae, S.-H. Yoon, T.-K. Kim, and K.-J. Yoon. Evidistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 608–619, 2021. 2
- [68] Z. Wang, Y. Lu, and L. Wang. Revisit event generation model: Self-supervised learning of event-to-video reconstruction with implicit neural representations. In *European Conference on Computer Vision*, pages 321–339. Springer, 2024. 1
- [69] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 2
- [70] W. Weng, Y. Zhang, and Z. Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 1
- [71] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 8
- [72] Z. Wu, M. Gehrig, Q. Lyu, X. Liu, and I. Gilitschenski. Leod: Label-efficient object detection for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16933–16943, 2024. 2
- [73] R. Xia, C. Zhao, M. Zheng, Z. Wu, Q. Sun, and Y. Tang. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21572–21581, 2023. 1, 3
- [74] R. Xie, C. Wang, W. Zeng, and Y. Wang. An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11240–11249, 2021. 2
- [75] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, and C. Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4968–4978, 2020. 2
- [76] M. Yan, Y. Zhang, S. Cai, S. Fan, X. Lin, Y. Dai, S. Shen, C. Wen, L. Xu, Y. Ma, et al. Reli11d: A comprehensive multimodal human motion dataset and method. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2250–2262, 2024. [2](#)
- [77] M. Yang, S.-C. Liu, and T. Delbruck. A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding. *IEEE Journal of Solid-State Circuits*, 50(9):2149–2160, 2015. [1](#)
- [78] W. Yang, J. Wu, J. Ma, L. Li, and G. Shi. Motion deblurring via spatial-temporal collaboration of frames and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6531–6539, 2024. [1](#)
- [79] Y. Yang, J. Han, J. Liang, I. Sato, and B. Shi. Learning event guided high dynamic range video reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13924–13934, 2023. [1](#)
- [80] Z. Yao and M. C. Chuah. Event-guided low-light video semantic segmentation. *arXiv preprint arXiv:2411.00639*, 2024. [1](#)
- [81] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. [4](#)
- [82] Y. Zhao, D. Rozumnyi, J. Song, O. Hilliges, M. Pollefeys, and M. R. Oswald. Human from blur: Human pose tracking from blurry images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14905–14915, 2023. [2](#)
- [83] X. Zheng and L. Wang. Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17448–17458, 2024. [1](#), [2](#)
- [84] J. Zhuang, Z. Wang, and Y. Gao. Semi-supervised video semantic segmentation with inter-frame feature reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3263–3271, 2022. [2](#)
- [85] S. Zou, C. Guo, X. Zuo, S. Wang, P. Wang, X. Hu, S. Chen, M. Gong, and L. Cheng. Eventhpe: Event-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10996–11005, 2021. [2](#)