

Article

EvTransPose: Towards Robust Human Pose Estimation via Event Camera

Jielun He , Zhaoyuan Zeng, Xiaopeng Li  and Cien Fan *

School of Electronic Information, Wuhan University, Wuhan 430072, China; eis-hjl@whu.edu.cn (J.H.); zhaoyuan.zeng@whu.edu.cn (Z.Z.); xiaopengli2014@whu.edu.cn (X.L.)

* Correspondence: fce@whu.edu.cn

Abstract: To overcome the interference caused by varying lighting conditions in human pose estimation (HPE), significant advancements have been made in event-based approaches. However, since event cameras are only sensitive to illumination changes, static human bodies often lead to motion ambiguity, making it challenging for existing methods to handle such cases effectively. Therefore, we propose EvTransPose, a novel framework that combines an hourglass module for global dependencies and a pyramid encoding module for local features. Specifically, a transformer for event-based HPE is adopted to capture the spatial relationships between human body parts. To emphasize the impact of high resolution on HPE tasks, this work designs the cascading hourglass architecture to compress and recover the resolution of feature maps frequently. On this basis, an intermediate-supervision constraint is incorporated to guide the network in aggregating sufficient features during the intermediate stages, which ensures better feature refinement and enhances overall performance. Furthermore, to facilitate a thorough evaluation of our method, we construct the first event-based HPE dataset with RGB reference images under diverse lighting conditions. Comprehensive experiments demonstrate that our proposed EvTransPose framework outperforms previous methods in multiple aspects.

Keywords: event camera; human pose estimation; intermediate supervision; transformer



Academic Editor: Valeri Mladenov

Received: 15 February 2025

Revised: 5 March 2025

Accepted: 7 March 2025

Published: 8 March 2025

Citation: He, J.; Zeng, Z.; Li, X.; Fan, C. EvTransPose: Towards Robust Human Pose Estimation via Event Camera. *Electronics* **2025**, *14*, 1078. <https://doi.org/10.3390/electronics14061078>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Human pose estimation (HPE) aims to localize human keypoints and plays a crucial role in various computer vision applications. Vision-based HPE has been extensively investigated in recent years, primarily adopting conventional sensors such as motion capture systems or RGB-D vision sensors [1–4]. However, approaches with these sensors are inherently limited by their imaging mechanisms, which pose challenges in capturing fine texture details of the target, especially under adverse lighting conditions such as darkness, backlighting, and overexposure. These limitations hinder the effectiveness of traditional sensor-based methods in scenarios where illumination is suboptimal.

In contrast to traditional frame-based cameras, event cameras, which asynchronously measure per-pixel brightness changes, offer advantages such as a high dynamic range (HDR, 120 dB), low power consumption, and high temporal resolution (up to 1 μ s) [5,6]. Recent work has also revealed the potential of event cameras in HPE, as their asynchronous imaging characteristic effectively mitigates the interference caused by lighting conditions and focuses on dynamic motion [7–10]. However, owing to the dependence on relative changes in brightness, event cameras struggle to capture events generated by a static human body under constant lighting conditions. During this time interval, the captured events

may correspond to multiple human poses, a phenomenon referred to as “motion ambiguity” in the context of event-based human pose estimation [11]. This ambiguity arises since static human body parts can exhibit multiple pose states, and the corresponding pose changes reflected by events may be indistinguishable.

A promising approach to overcoming the motion ambiguity issue, without introducing additional temporal information, is to supplement and predict the static local part of the target by modeling global dependencies with sufficient information from other parts [12–15]. The validity of this idea has been verified in MAE [16], and the Vision Transformer (ViT) [17] has made a key contribution to this process. Furthermore, local features cannot be ignored in event-based HPE since adjacent pixels still contain motion information about the target, even though they lack texture details. Based on the above analysis, there are three key challenges in achieving robust and powerful performance in event-based HPE:

- How to make full use of global dependencies to accurately predict the human joint coordinates in the presence of motion ambiguity.
- How to ensure the effective extraction of local features on this basis for high accuracy in normal prediction.
- How to maintain the consistency of global dependencies and local features to avoid the negative impact of their conflict on the estimation results.

Recent research in related fields has provided valuable insights that can help address these issues. For instance, studies on cognitive computing-based flexible rectifiers for flow state prediction [18] have demonstrated the potential of adaptive algorithms in dynamic and uncertain environments, which aligns with the challenges faced in event-based HPE. In a similar vein, research on voltage-driven oscillating liquid-flow robots [19] has explored how dynamic control can be applied to fluid systems to achieve precise movement. Although the aforementioned work focused on mechanical systems, the underlying concepts of dynamic control share similarities with the challenges we address in our framework, particularly when handling static body poses under varying lighting conditions.

To address the above challenges, we propose EvTransPose, a novel model with an hourglass module for capturing global dependencies and a pyramid encoding module for encoding local features to realize event-based HPE. A transformer-based hourglass module is proposed to effectively integrate global dependencies by leveraging the self-attention mechanism. In detail, the pyramid encoding module is adopted to efficiently collect and flatten the local feature maps as input to the self-attention module at different scales, and the deconvolution decoding module is utilized to restore the output to heatmaps with the same resolution as the original input. To preserve the comprehensive perception of HPE during feature extraction and maintain consistency between local and global features, our model employs a cascade of hourglass modules, which facilitates the iterative process of top-down compression and bottom-up recovery during inference. In summary, the main contributions of our work are as follows:

1. We adopt an attention mechanism for event-based HPE to capture the spatial relationships between human body parts, which can efficiently tackle the challenge of motion ambiguity.
2. We introduce an intermediate-supervision constraint to aggregate features from a more balanced perspective, facilitating the coordination of local and global features. Additionally, a cascading hourglass architecture is proposed to make intermediate features learnable and conditioned by loss functions.
3. We design a hybrid imaging system capable of capturing event streams and RGB images simultaneously and build the first HPE dataset based on events with RGB reference images in a variety of lighting conditions, providing strong support for subsequent research.

The remainder of this paper is organized as follows. Section 2 reviews previous works related to our proposed method, including frame-based HPE methods and event-based ones. Our proposed framework, including the HPE network, the cascading hourglass architecture, and relevant loss functions, is described in Section 3, and the event-based HPE dataset is presented in Section 4. Section 5 reports qualitative and quantitative comparisons with several state-of-the-art methods, followed by ablation studies. Finally, the conclusions are given in Section 6.

2. Related Works

Although there are also many works based on wearable sensors in HPE that collect data through Inertial Measurement Units (IMUs) [20], accelerometers [21], and gyroscopes [22] and output results through a Kalman filter [23,24], they are inherently constrained by the need for various wearable sensors and the location drift issue [25,26]. Similar challenges are also present in the field of human–robot collaboration (HRC) [27]. Consequently, this study places more emphasis on vision-based HPE.

2.1. Frame-Based Human Pose Estimation

Previous HPE methods consisted of regression-based methods and heatmap-based methods. Earlier studies on HPE mainly focused on regression since HPE is naturally regarded as a regression problem. Typical methods generally extracted features from input by designing hand-crafted features such as SIFT [28] and HOG [29]. Although some authors attempted to develop a hand-designed model for HPE or an aligning mechanism to better learn the structure-aware information [30,31], regression-based methods are not as accurate as heatmap-based ones [32,33]. The main reason for this is that a heatmap combined with the original image can retain informative vision features, and the task is transformed into a segmentation problem, which operates at the pixel level, rather than a highly nonlinear coordinate regression problem. Recent heatmap-based methods tend to stack deeper architectures or preserve multi-resolution feature maps to enhance the quality of heatmaps [34–37]. However, due to the inherent locality of convolution, CNNs reveal their limitation in capturing and modeling constraint relationships between keypoints, which is crucial for HPE.

Due to its attention layers, the transformer architecture [38] has a natural advantage over CNNs in terms of capturing interactions between any pairwise locations [39–43]. In the context of HPE, transformers can be well exploited to model global dependencies, thus alleviating the motion ambiguity issue regardless of whether directly regressing [44] or predicting heatmaps [45,46].

2.2. Event-Based Human Pose Estimation

In event-based vision, there have been very few investigations addressing human pose estimation. Existing methods use different representations of events, such as event count images (ECIs), voxels, point clouds, and their variants. Represented by TORE [47], voxels preserve event characteristics and informative temporal features. However, events with similar time stamps may be placed in different channels, resulting in significant time disparities within the same channel, which destroys the coherence of the space-time structure to some extent. In contrast with ECIs, by making use of the natural properties of events as synchronous signals with high temporal resolution, point clouds offer lower latency and reduced computational overhead [48].

One of the most effective and general strategies is to represent the event data as synchronized event frames so as to directly use the mature frame-based HPE network. Recent methods have demonstrated the advantages of this strategy and have built event-based

real-world datasets [49] or simulation datasets [50], but they are still scarce. Furthermore, they have highlighted the challenge of motion ambiguity and its limitation in event-based HPE tasks. However, recent works have focused on incorporating additional information sources, like grayscale images [51,52] or novel representations of events [47,48], which complicates both data acquisition and forward networks.

3. The Proposed Method

In this section, we first present the detailed design of our EvTransPose framework in Section 3.1, followed by the intermediate-supervision constraint in Section 3.2.

3.1. Network Architecture

As illustrated in Figure 1, our proposed EvTransPose mainly consists of M_1 cascading hourglass modules, which are responsible for extracting semantic information from the input event frames. It is composed of four components: one pyramid encoding module, M_2 cascading self-attention modules, one deconvolution decoding module, and one pose estimation head module. Next, we elaborate on the details of the composition of the hourglass module in order of precedence.

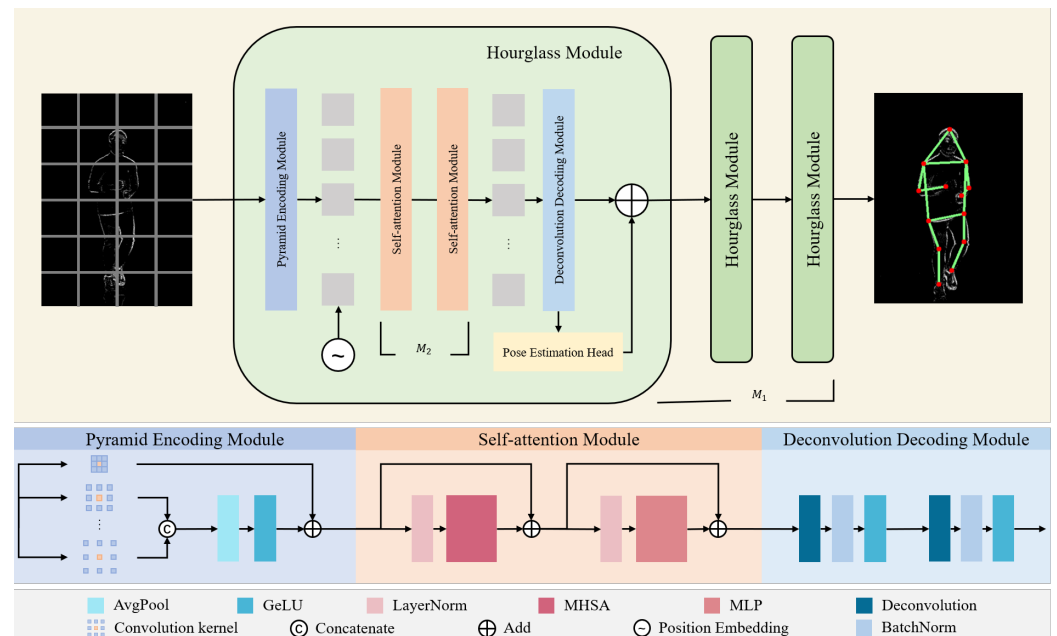


Figure 1. Schematic illustration of the proposed EvTransPose framework. Event frames flow through the M_1 cascading hourglass modules to produce the expected HPE results. In detail, the pyramid encoding module compresses the resolution of the input and encodes feature maps to obtain the block embeddings. Next, these embeddings are sent to the M_2 cascading self-attention modules, focusing on integrating global dependencies and generating feature maps. Then, the deconvolution decoding module recovers the resolution of feature maps, and the pose estimation head module predicts the human joint coordinates.

3.1.1. Pyramid Encoding Module

The sizes of human bodies in images vary significantly, leading to different spatial distributions of individuals within the frame, which makes it difficult to extract local features using convolutions at a fixed scale. Therefore, a multi-parallel pyramid structure convolution, as shown in Figure 2, is utilized to expand the receptive field while maintaining a low parameter count, which consists of two main branches. In the first branch, input images flow through a convolution with a kernel size of $s \times s$ and a step size of $s \times s$ for S -fold downsampling. Specifically, as input, a feature frame $F_{in} \in \mathbb{R}^{C \times W \times H}$ is embedded by

the first branch into feature vectors, i.e., $F_1 \in \mathbb{R}^{(s^2C) \times (\frac{W}{s}) \times (\frac{H}{s})}$, where s is the downsampling ratio and C is the channel dimension. The other branch flows through the multi-level feature extraction (MFE) module which includes four parallel convolutions with a kernel size of 3×3 and dilation rates of 1, 2, 3, and 4, respectively, resulting in feature vectors $F_k \in \mathbb{R}^{C \times W \times H}$, where $k = 1, 2, 3, 4$. By concatenating these feature vectors together, one can obtain features with comprehensive information on multi-scale human body parts, followed by an AvgPool for S -fold downsampling. To ensure that the intermediate vector has the same number of channels as F_1 , a 1×1 convolution is subsequently applied. Finally, the result is added to F_1 and flattened, yielding the expected embeddings $F \in \mathbb{R}^{N_{emb} \times d}$, where $N_{emb} = \frac{WH}{s^2}$ and $d = s^2C$. Compared with single-scale convolutional blocks, our pyramid encoding module maps data into the latent space to obtain more robust feature representations, thereby allowing for more accurate human joint coordinate localization under challenging conditions.

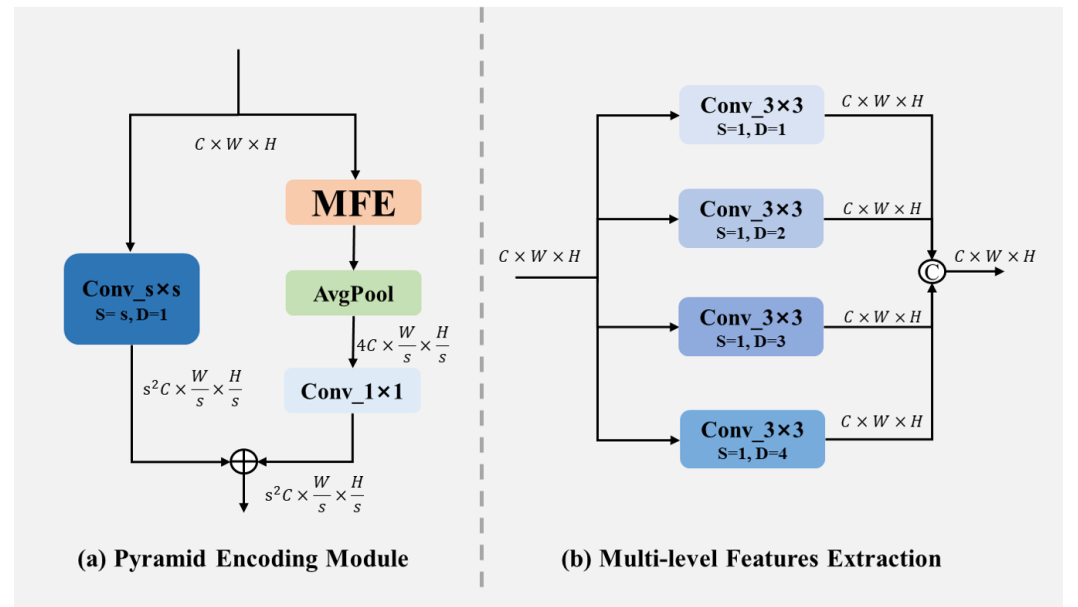


Figure 2. (a) Architecture of the pyramid encoding module. (b) Architecture of the MFE module, where S and D denote the number of strides and dilation rates, respectively.

3.1.2. Self-Attention Module

To fully leverage global dependencies for accurate HPE prediction amidst motion ambiguity, self-attention (SA) is incorporated into the hourglass modules. For the initial feature $F \in \mathbb{R}^{N_{emb} \times d}$, self-attention is formulated as

$$SA(F) = \text{Softmax}\left(\frac{FW_Q(FW_K)^T}{\sqrt{d}}\right)FW_V, \quad (1)$$

where the parameters of the fully connected layer ($W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$) are exploited to calculate the queries Q , keys K , and values V . In addition, in order to focus on features with multi-scale spatial information from different perspectives, multi-head self-attention (MHSA) is adopted, and the process is expressed as

$$MHSA(F) = \text{CONCAT}(SA_i(F))|_{i=1}^{N_{head}} \times W_O. \quad (2)$$

In this context, N_{head} denotes the number of SA heads, and $W_O \in \mathbb{R}^{(N_{head}d) \times d}$ represents the parameters of the fully connected layer for transforming the different outputs of each SA head to the same size. Overall, our self-attention module is expressed as

$$\widehat{F}_{l-1} = F_{l-1} + \text{MHSA}(\text{LN}(F_{l-1})), \quad (3)$$

$$F_l = \widehat{F}_{l-1} + \text{MLP}\left(\text{LN}\left(\widehat{F}_{l-1}\right)\right), \quad (4)$$

where LN and MLP are layer normalization and a multi-layer perceptron, respectively, and l indicates that the feature is computed by the l th self-attention module in each hourglass module.

3.1.3. Deconvolution Decoding Module

To ensure the continuity of the hourglass module, the features need to be recovered to the input size $\mathbb{R}^{C \times W \times H}$. In the original transformer framework, embeddings are treated as independent entities, with each embedding vector directly reshaped during decoding to recover the original resolution. However, this approach disrupts spatial associations between features across different channels, thereby compromising the global dependencies captured by the attention mechanism. To mitigate this issue, deconvolution is adopted instead of direct reshaping to preserve the global dependencies learned during the self-attention module. This process can be formulated as

$$F_{out} = \text{Deconv}(F_{M_2}). \quad (5)$$

3.1.4. Pose Estimation Head Module

In order to validate the superiority of our network structure, a 1×1 convolution layer is employed to generate the localization heatmaps for the keypoints, e.g.,

$$Y = \text{Conv}_{1 \times 1}(F_{out}), \quad (6)$$

where $Y \in \mathbb{R}^{N_{joint} \times W \times H}$ and N_{joint} denotes the number of HPE keypoints.

3.2. Intermediate-Supervision Constraint

By introducing richer intermediate features, the network is encouraged to aggregate features from a more balanced perspective, facilitating the coordination of local and global features. In a network utilizing a single hourglass module, the embeddings can capture both local features and global dependencies only after passing through at least one pyramid encoding module and one self-attention module. However, these embeddings are not suitable as intermediate features for constraining the network against the ground truth unless they are decoded. Therefore, to preserve the overall consistency of features and alleviate problems of anatomic violation, the cascading hourglass architecture divides the continuous self-attention modules across individual hourglass modules, which ensures that feature extraction is learnable and conditioned by the associated loss functions. Moreover, in practice, it is essential to carefully design the connections between cascading hourglass modules.

3.2.1. Loss Functions

To enhance the inductive bias for lack of locality in transformers, multiple units are connected in a sequential manner, iterating through a top-down compression and bottom-up recovery process. Consequently, it is imperative to introduce a series of loss functions to constrain our proposed EvTransPose method and achieve the desired outcomes. To be specific, mean squared error (MSE) loss is employed to narrow the difference between the

generated heatmaps and the ground-truth ones during the training process. For the k th hourglass module, the loss is formulated as

$$\begin{aligned} L_{\text{hpe}}^k(Y, GT) &= \sum_{i=1}^{N_{\text{joint}}} w_i L_{\text{MSE}}(Y_i, GT_i) \\ &= \sum_{i=1}^{N_{\text{joint}}} \frac{w_i}{N_{\text{pix}}} \sum_{j=1}^{N_{\text{pix}}} (Y_{i,j} - GT_{i,j})^2, \end{aligned} \quad (7)$$

where Y and GT denote the output of the pose estimation head module and the ground truth, respectively; N_{joint} and N_{pix} are the number of HPE keypoints and the number of pixels in the heatmaps, respectively; and w_i represents the weight coefficient of the skeleton keypoints. For the cascading hourglass architecture, it is imperative to design a set of balanced loss weights. The network will overvalue the intermediate features and cause severe degradation in the subsequent modules if too much emphasis is placed on the early stage. However, intermediate supervision will weaken its constraints if too much emphasis is placed on the later stage. Thus, we design a set of exponential loss weights for each stage:

$$L_{\text{hpe}}(Y, GT) = \sum_{k=1}^{M_1} \lambda_k L_{\text{hpe}}^k(Y, GT). \quad (8)$$

In this context, M_1 is the total number of hourglass modules, and the loss weights λ_k can be formulated as

$$\lambda_k = \eta^{k-M_1}, \quad (9)$$

where η is an attenuation coefficient to control the decay rate of the function.

3.2.2. Inter-Module Connection

The spatial information embedded in features is vital for accurately positioning the coordinates of the skeleton keypoints in HPE tasks. Therefore, the features generated by each hourglass module are interconnected through skip connections to coherently retain shallow spatial information. Moreover, these interconnections help prevent vanishing gradients caused by the deepening of the network structure [50] and stabilize the training process. The concrete selection of the skip connections is detailed in Section 5.

4. Dataset

Current studies on event-based HPE are still in their early stages, and an important factor restricting development is the limited availability of datasets. At present, the only real-world event dataset publicly available in this field is DHP19 [49]. To support our experiments and facilitate future research, we designed a hybrid imaging system capable of collecting event streams alongside simultaneous RGB images, creating the first event-based HPE dataset with RGB frames as references under various lighting conditions, namely the Challenging Human Pose (CHP) dataset.

4.1. Data Acquisition

As displayed in Figure 3, the CHP dataset was recorded by a hybrid vision camera system equipped with a beamsplitter [53], which connects an event camera, i.e., SilkyEvCam, to capture concurrent event streams, with a traditional RGB camera, i.e., FLIR BFS-U3-32S4, to collect matched RGB images. Notably, aligning the two different modes in both time and space is crucial.

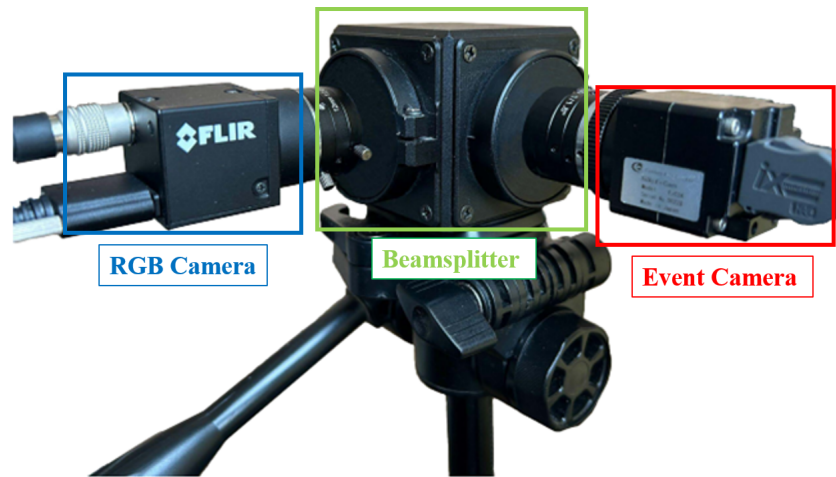


Figure 3. Implementation of our hybrid vision camera system.

4.1.1. Time Synchronization

Our system adopts low-latency hard synchronization and strictly aligns the two sets of acquired data with a shared set of timestamps [54]. In detail, the RGB camera is the master that sends triggers while collecting images, and the event camera is the slave that captures events and records timestamps while receiving triggers. The sync signal is transmitted between the sync ports of the two cameras through a high-speed data line, ensuring that both cameras share the same set of timestamps, which indicate the time of data acquisition.

4.1.2. Space Calibration

The input light is split into two mirror images captured by the two cameras; thus, in addition to flipping the images left to right, we need to calibrate the imaging differences caused by the cameras' parameter matrices [55,56]. Concretely, the process of casting the three-dimensional space coordinates Z_{real} onto a two-dimensional plane (camera imaging) can be described as

$$\begin{pmatrix} Z \\ 1 \end{pmatrix} = P_{in}(R, t) \begin{pmatrix} Z_{real} \\ 1 \end{pmatrix}, \quad (10)$$

where $P_{in} \in \mathbb{R}^{3 \times 3}$ defines the internal parameter matrix of the cameras, which is invertible, and $R \in \mathbb{R}^{3 \times 3}$ and $t \in \mathbb{R}^{3 \times 1}$ are the rotation matrix and translation vector of the cameras, respectively, constituting the external parameter matrix of the camera. The effect of the translation vector t can be approximately ignored when the imaging scene is deep enough. Then, by defining the rotation calibration matrix R_{trans} , we obtain

$$\begin{pmatrix} Z_E \\ 1 \end{pmatrix} = P_E R_{Trans} P_{RGB}^{-1} \begin{pmatrix} Z_{RGB} \\ 1 \end{pmatrix}. \quad (11)$$

During execution, P_E and P_{RGB} are the internal parameter matrices of the event camera and RGB camera, respectively, and Z_E and Z_{RGB} are the two-dimensional coordinates of the same target captured by the event camera and RGB camera, respectively. By collecting several pairs of targets captured after time synchronization, our system can use least squares to obtain the rotation calibration matrix R_{trans} , based on which two different modes can be accurately aligned at the spatial level.

4.1.3. Space Event Representation

We choose to process event frames instead of raw event streams for several reasons: Event frames aggregate temporal information, making it easier to capture spatiotemporal dependencies. This approach is computationally more efficient and allows us to leverage

existing deep learning architectures, such as CNNs and transformers, which are designed for image-like data. Ultimately, event frames strike a balance between performance and computational efficiency.

Specifically, for the i -th event frame, which corresponds to an RGB image frame with timestamp t_i , the event frame pixels are obtained by accumulating the events that occurred within the time window $[t_{i-1}, t_i)$. Let the event stream during this time period be $\varepsilon_i = [e_j]_{j=1}^{N_i}$. The pixel $E(x, y)$ at coordinates (x, y) in the event frame can be defined as follows:

$$E(x, y) = \frac{E'(x, y)}{\max(E'(x, y))} \times 255, \quad (12)$$

$$E'(x, y) = \sum_{j=1}^{N_i} \delta(x - x_j, y - y_j), \quad (13)$$

where δ is the Dirac function, and x_j and y_j define the coordinates (x, y) of event e_j , respectively.

4.2. Dataset Description

The CHP dataset contains 12 predefined movements recorded from eight subjects, including 192 pairs of event and RGB image sequences, totaling 28,800 frames. The dataset is partitioned into training and testing subsets at a ratio of 3:1. The event streams and RGB images are processed into frames with a resolution of 640×480 under the same field of view for training. As shown in Figure 4, underexposed images, normally exposed images, overexposed images, and synchronized event frames are depicted from left to right.

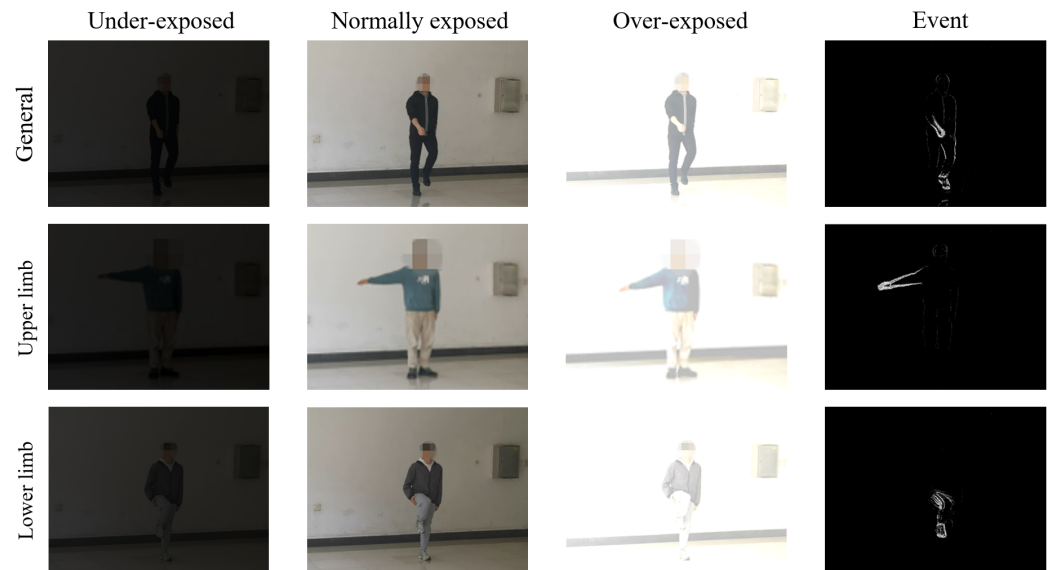


Figure 4. Samples from our proposed CHP dataset, depicting underexposed images, normally exposed images, overexposed images, and synchronized event frames, from left to right, and composed of three movement groups: general movement, upper-limb movement, and lower-limb movement from top to bottom.

Based on the DHP19 dataset, we redefine certain movements that are difficult to classify clearly and reorganize the 12 movements into three categories: general movements (3), upper-limb movements (5), and lower-limb movements (4). The general movements provide abundant event-related information, representing a general scenario for event-based HPE. In contrast, upper-limb or lower-limb movements are often subject to motion ambiguity.

ity, representing a more challenging scenario. This restructuring enables us to use the CHP dataset to more effectively assess the network's ability to extract local features and infer local posture from global dependencies, particularly in the presence of adverse lighting conditions and motion ambiguity. Specifically, to enhance the robustness of the model against variations in human pose and illumination, we applied various augmentation techniques, including random scaling, rotation, flipping, and noise injection, to simulate different lighting and environmental conditions.

5. Experimental Results and Discussion

5.1. Experimental Settings

5.1.1. Evaluation Metrics

Consistent with prior evaluation metrics [49,52,57], we adopted the mean per-joint position error (MPJPE), which is a widely used metric in HPE and can be calculated in both 2D and 3D spaces (denoted as $MPJPE_{2D}$ (in pixels) and $MPJPE_{3D}$ (in mm), respectively). While MPJPE provides an intuitive measure of the distance between the ground truth and the predicted joints, it is subject to bias due to the perspective projection. To mitigate this, we introduced the percentage of correct keypoints (PCK), a head-normalized metric that is independent of camera intrinsic parameters and imaging conditions, offering a more robust evaluation.

5.1.2. Implementation Details

We implemented our proposed EvTransPose framework in PyTorch 1.5.1 and used Adam [58] as the optimizer. Model training was executed on four NVIDIA GeForce RTX 1080 Ti GPUs for 210 epochs. The initial learning rate was set to 0.0005. The learning rate was reduced linearly by a factor of 10 at the 170th and 200th epochs. In addition, a linear warm-up with an initial value of 0.1% was applied for the first 500 iterations.

5.2. Evaluations on the DHP19 Dataset

For the single-modal dataset DHP19, to compare with existing relevant state-of-art HPE methods, we selected eleven competitive methods, categorized as CNN-based methods (SHG [34], PoseRes [36], HRNet [37], DHP19 [49], TORE [47], and DGCNN [59]), transformer-based methods (HRFormer [60], ViTPose [45], VMST-Net [61], and PointTransformer [62]), and PointNet-based method (PointNet [63]).

The quantitative results are shown in Table 1. In normal scenes, our method outperformed the baseline method, ViTPose, in $MPJPE_{2D}$ and $MPJPE_{3D}$ by 0.26 pixels and 7.93 mm, respectively. Notably, the transformer-based methods consistently performed better than the others, highlighting the advantages of the self-attention mechanism in capturing the global dependencies in HPE. Due to the FIFO sequence at each pixel, which incorporates abundant temporal information, TORE mitigated motion ambiguity and delivered highly accurate estimations. However, in contrast to the voxel-based or point cloud-based methods, our approach predicted more precise human joint coordinates from a single frame. This can be attributed to our cascading hourglass architecture, which iteratively compresses and recovers informative feature maps to facilitate the effective integration of both local and global features, thereby addressing motion ambiguity.

The qualitative results presented in Figure 5 further validate the effectiveness of our method. While all algorithms demonstrated commendable performance in general movement, our method was the most effective. In scenarios involving typical motion ambiguity, the transformer-based approaches provided relatively accurate predictions with sparse event information compared to the CNN-based methods. For instance, the keypoint predictions for the right foot in the boxed region exhibit noticeable differences among the

various methods. Thanks to our novel pyramid encoding module and cascading hourglass architecture, the proposed method preserves the consistency between global dependencies and local features, enabling more precise localization of human keypoints. Therefore, it outperformed other transformer-based approaches in terms of accuracy and refinement.

Table 1. Quantitative comparison of different methods on the DHP19 dataset.

Representations	Frameworks	Methods	MPJPE _{2D} ↓	MPJPE _{3D} ↓
Frame	CNN	SHG	8.29	92.16
	CNN	PoseRes	6.25	62.26
	CNN	HRNet	6.14	64.63
	CNN	DHP19	7.67	79.87
	Transformer	HRFormer	6.21	65.89
	Transformer	ViTPose	<u>5.34</u>	58.32
	Transformer	EvTransPose	5.08	50.39
Voxel	CNN	TORE	5.44	<u>56.46</u>
	Transformer	VMST	6.45	73.07
Point Cloud	CNN	DGCNN	6.83	77.32
	PointNet	PointNet	7.29	82.46
	Transformer	PointTransformer	6.46	73.37

The best results are highlighted in **bold**, and the second-best results are underlined.

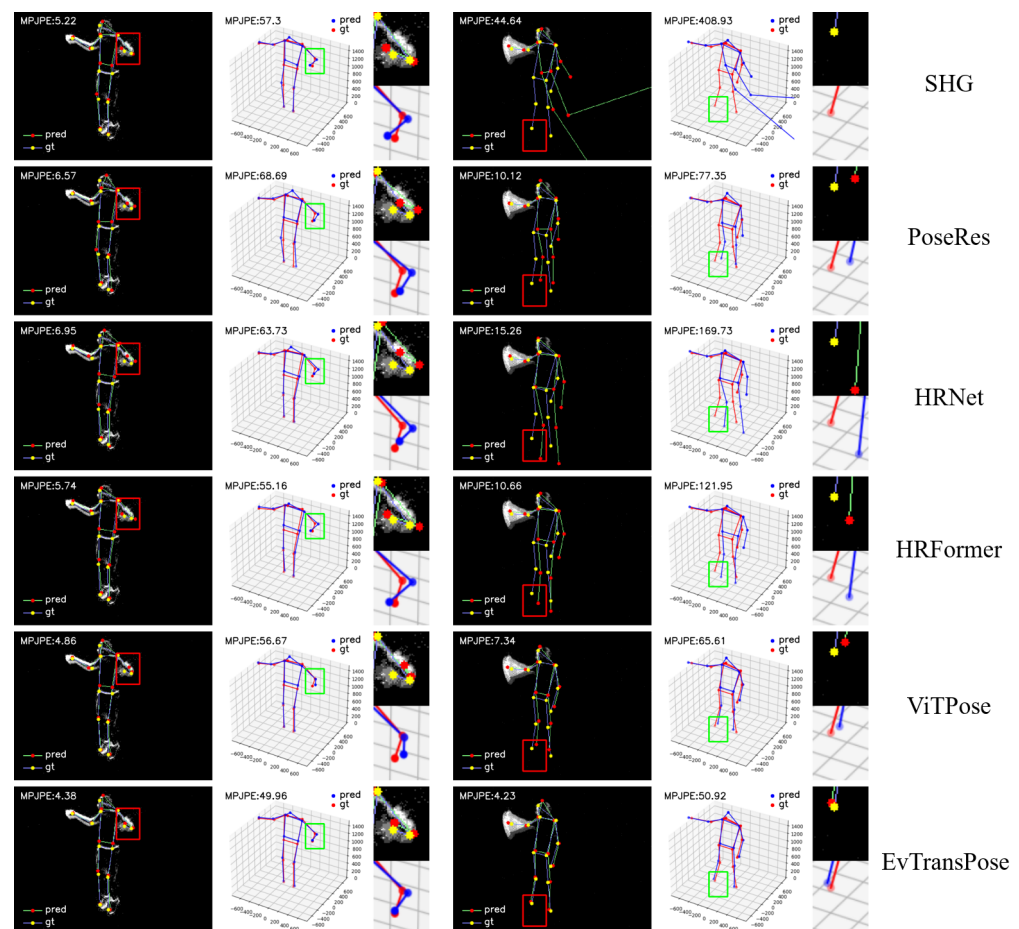


Figure 5. Qualitative results and visualization of 3D pose estimation. The left two columns display a jumping movement, which falls under general movement and generates abundant event information, while the right two columns display a waving movement, which falls under upper-limb movement and leads to motion ambiguity.

5.3. Evaluations on the CHP Dataset

Since our method is based on events, which are represented by event frames, we selected the frame-based methods discussed earlier as baselines to assess the performance of our approach. Notably, all methods were retrained on our CHP dataset to ensure consistency and validity in the evaluations.

5.3.1. Quantitative Evaluation

As shown in Table 2, the proposed EvTransPose achieved superior performance across all metrics. In particular, owing to its excellent ability to extract local features, our method demonstrated significant improvements over state-of-the-art frame-based methods, especially under more stringent thresholds. To further demonstrate the robustness of our method, Table 3 presents comparative results across various movements in the CHP dataset. The results indicate that, for more complex movements, the transformer-based approaches were more effective, and the CNN-based approaches exhibited noticeable performance gaps between general movements and upper- or lower-limb movements due to their neglect of global dependencies. HRNet and ViTPose experienced performance drops of 25.44% and 22.31% in terms of PCK@0.1, respectively, while our method maintained a much smaller decrease of only 14.74%. All the results underscore the robustness and broad applicability of the proposed method across diverse movement scenarios.

Table 2. Quantitative comparison of different frame-based methods on the CHP dataset.

	SHG	PoseRes	HRNet	HRFormer	ViTPose	EvTransPose
MPJPE _{2D} ↓	89.25	18.20	19.45	19.83	<u>17.88</u>	17.34
PCK@0.1↑	19.45	57.08	55.85	56.12	<u>62.94</u>	67.24
PCK@0.3↑	46.46	<u>93.07</u>	92.50	92.36	92.81	94.10

PCK is an accuracy out of 1, where 0.1 and 0.3 are the threshold values. The best results are highlighted in **bold**, and the second-best results are underlined.

Table 3. Comparison of different frame-based methods across various movements in the CHP dataset.

Movement		PCK@0.1↑			
		SHG	HRNet	ViTPose	EvTransPose
General Movement	1	21.38	66.62	76.23	<u>75.54</u>
	2	28.00	78.77	<u>81.15</u>	81.38
	3	40.62	76.77	81.62	<u>78.08</u>
Upper/Lower-Limb Movement	4	21.38	49.08	<u>59.31</u>	64.54
	5	17.38	37.23	<u>49.85</u>	62.08
	6	24.31	43.69	<u>53.77</u>	64.08
	7	10.15	44.00	<u>52.15</u>	63.76
	8	10.46	54.31	58.85	<u>55.46</u>
	9	15.69	60.77	<u>68.08</u>	73.00
	10	10.15	43.23	<u>52.38</u>	62.08
	11	18.92	57.23	<u>59.00</u>	67.31
	12	14.92	55.86	62.85	<u>62.23</u>
Mean		19.45	55.85	<u>62.94</u>	67.24

The best results are highlighted in **bold**, and the second-best results are underlined.

5.3.2. Qualitative Evaluation

For the qualitative evaluation, we present a visual comparison of various methods on the CHP dataset in Figure 6, which includes event frames captured under different lighting conditions. In the comparison, we provide the corresponding RGB images in the leftmost

column. Regardless of the illumination conditions, our proposed framework consistently delivered promising estimations. It fully demonstrated the substantial advantages of events over RGB images, particularly in situations where texture details were severely degraded by poorly exposed regions. Furthermore, taking the images in the third row as an example, the transformer-based approaches still achieved strong performance, even with sparse event information, whereas the CNN-based approaches tended to fail in scenarios with motion ambiguity. This can be attributed to the self-attention module, with our designed intermediate-supervision constraint optimizing its performance to generate more visually pleasing and accurate estimations.

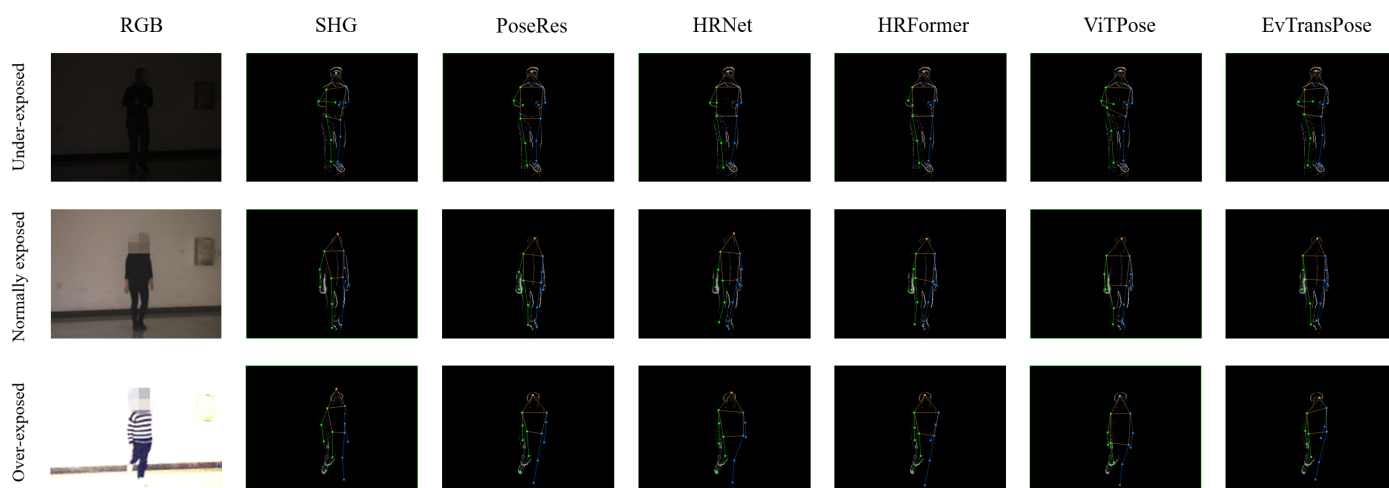


Figure 6. Qualitative results on the CHP dataset, where our proposed EvTransPose framework is compared with different frame-based methods. Rows from (top) to (bottom): underexposed images, normally exposed images, and overexposed images.

5.4. Ablation Study

We conducted an ablation study on each key module of our proposed EvTransPose. Specifically, we evaluated the effect of the pyramid encoding module, the effectiveness of the cascading hourglass architecture, the importance of intermediate supervision, and conducted cross-dataset testing.

5.4.1. Importance of Pyramid Encoding Module

To demonstrate the effect of the pyramid encoding module, we compared it with other widely used encoding methods for transformers, including direct token reserialization in ViT [17] and one-step block convolution in ViTPose [45]. As illustrated in Table 4, the proposed pyramid encoding module effectively harnessed parallel multi-scale convolutions to maximize the receptive field with limited parameters, resulting in a superior ability to extract local features.

Table 4. The performance of EvTransPose using different encoding methods on DHP19.

	MPJPE _{3D} ↓	MPJPE _{2D} ↓	PCK@0.1↑	PCK@0.3↑
ViT	61.93	6.80	66.93	98.45
ViTPose	<u>54.79</u>	<u>5.46</u>	<u>73.47</u>	<u>98.88</u>
Ours	50.39	5.08	76.84	99.20

The best results are highlighted in **bold**, and the second-best results are underlined.

5.4.2. Importance of Cascading Hourglass Architecture

We further validated the effectiveness of the cascading hourglass architecture, focusing on its contribution and the impact of different cascading strategies. In our experimental

configuration, the model is cascaded in three distinct manners: by connecting without skip connections, by concatenating, and by adding up. Here, M_1 and M_2 denote the numbers of cascaded hourglass modules and cascaded self-attention modules, respectively, while maintaining a total of twelve self-attention modules. The results in Table 5 indicate that an increasing number of cascaded hourglass modules led to performance degradation in models without skip connections due to gradient vanishing issues. In contrast, models with skip connections showed consistent improvements in prediction accuracy, among which adding up yielded superior results compared to concatenating across all evaluation metrics. This can be attributed to the ability of the adding strategy to preserve shallow spatial information, which is critical for the accurate localization of keypoints.

Table 5. The performance of EvTransPose with different cascading architectures on DHP19.

Methods	M_1	M_2	MPJPE _{3D} ↓	MPJPE _{2D} ↓	PCK@0.1↑	PCK@0.3↑
-	1	12	58.25	5.78	69.16	98.95
Without Skip Connection	2	6	60.06	5.94	67.55	98.81
	3	4	60.86	6.00	67.22	98.70
	4	3	346.23	29.50	4.03	13.60
	6	2	344.24	29.31	4.06	31.27
Concatenate	2	6	58.67	5.80	69.12	95.36
	3	4	55.42	5.52	72.51	99.00
	4	3	55.64	5.54	72.55	98.95
	6	2	<u>53.17</u>	<u>5.33</u>	<u>74.65</u>	99.06
Add (Ours)	2	6	57.31	5.69	70.22	98.90
	3	4	54.35	5.42	73.62	99.05
	4	3	54.00	5.39	73.78	<u>99.08</u>
	6	2	51.88	5.21	76.00	99.14

The best results are highlighted in **bold**, and the second-best results are underlined.

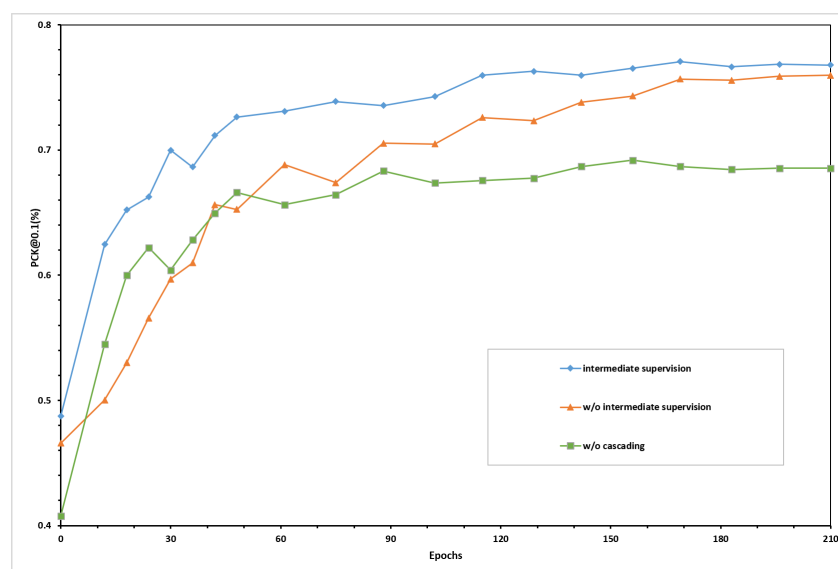
5.4.3. Analysis of Intermediate Supervision

To investigate the contribution of the added intermediate-supervision constraint and the influence of different loss weights, we conducted a comprehensive ablation study on DHP19 and present the results in Table 6 and Figure 7. During execution, η was an attenuation coefficient for the loss weights λ_k , e.g., $\lambda_k = 1$ for constant weights, $\lambda_k = \frac{k+\eta}{M_1+\eta}$ for linear weights, $\lambda_k = \log_{\eta M_1}(\eta k)$ for logarithmic weights, and $\lambda_k = \eta^{k-M_1}$ for exponential weights, where $k = 1, 2, \dots, M_1$. This observation validates that introducing intermediate supervision was beneficial for improving accuracy and accelerating the convergence of the network, highlighting the importance of maintaining an appropriate balance in the loss weights. When assigned an excessively high weight in the early stages, the network tended to overly focus on providing accurate pose estimation, potentially neglecting the extraction of the deep features and semantic information from the graphics. Conversely, intermediate supervision weakened and lost its competitiveness when assigned an excessive weight in the later stages. According to the experiments, adopting exponential loss weights, where $\eta = 10$, yielded the optimal balance, thereby enhancing the performance of EvTransPose.

Table 6. The performance of EvTransPose with different loss weights for intermediate supervision on DHP19.

Weights	None	Constant	Linear		Logarithmic		Exponential			
η	-	-	-1	1	1	2	2	5	10	20
MPJPE _{3D} *↓	-	<u>56.67</u>	58.36	59.54	59.49	58.75	56.20	84.19	114.39	125.38
PCK@0.1 *↑	-	<u>71.57</u>	69.27	69.42	69.33	69.02	71.87	53.95	34.60	30.35
PCK@0.3 *↑	-	98.91	98.89	98.49	98.55	98.88	<u>98.89</u>	95.30	91.67	90.23
MPJPE _{3D} ↓	51.88	55.97	57.40	58.32	58.10	57.93	52.72	<u>51.29</u>	50.39	51.63
PCK@0.1↑	76.00	72.18	70.06	70.79	70.86	69.79	75.31	76.27	76.84	<u>76.42</u>
PCK@0.3↑	99.14	98.92	98.95	98.55	98.61	98.91	99.05	99.16	99.20	<u>99.15</u>

The best results are highlighted in **bold**, and the second-best results are underlined. * denotes the evaluation metrics calculated with the intermediate results of the third hourglass module.

**Figure 7.** Convergence curves of EvTransPose with different strategies on DHP19, where “w/o intermediate supervision” indicates intermediate supervision was excluded during training, and “w/o cascading” indicates that the cascading hourglass architecture was not employed.

5.4.4. Cross-Dataset Testing

To assess performance degradation across diverse datasets, we conducted a comprehensive ablation study in which the model was trained on the source domain dataset DHP19 after data enhancement, and the trained model was tested on the target domain dataset CHP. As shown in Table 7, based on the network’s generalization capability, performance in the target domain was not impressive, with noticeable declines across various metrics. This can be attributed to significant differences between datasets, such as variations in movements and subjects. In addition to differences in annotation locations, the disparity in data distribution, known as domain shift, was a primary cause of the model’s poor performance on other datasets. In the field of event-based human pose estimation, this is manifested in factors such as event sparsity, noise intensity, and distribution. To address this issue, it may be necessary to further incorporate Domain Adaptation (DA) techniques.

Table 7. Results of cross-dataset testing.

Source Domain	MPJPE _{2D} ↓	PCK@0.1↑	PCK@0.3↑
DHP19	23.61	39.79	89.91
CHP	17.34	67.24	94.10

6. Conclusions

In this paper, we propose a novel neural network named EvTransPose to localize human keypoints of interest under challenging lighting conditions using a single event frame. In particular, an attention mechanism for event-based HPE is incorporated into the HPE pipeline to address motion ambiguity by leveraging the spatial dependencies between human body parts. To maintain a comprehensive perception of HPE and the consistency of local and global features, we introduce an intermediate-supervision constraint within the cascading architecture, which ensures that feature extraction is learnable and conditioned on the related loss functions. In addition, we construct a new HPE dataset, i.e., CHP, based on events with synchronous RGB images across diverse lighting scenarios. Experimental results, both quantitative and qualitative, show the effectiveness of our proposed EvTransPose.

By leveraging the sparse event information generated by immobile limbs to infer overall posture, our proposed EvTransPose alleviates motion ambiguity to a great extent. However, pose estimation performance still degrades notably under extreme cases where the entire body remains motionless. Future work will focus on incorporating additional signals like prior coordinates or heatmaps, fusing RGB information, and implementing spatiotemporal architectures. To fully address extreme situations, a new challenge may arise in that this approach makes the network more reliant on contextual information and requires maintaining a balance between performance and computational resources.

In future work, we will also explore methods for reducing the number of network parameters and decreasing computational complexity. By exploring lightweight variations such as shallow networks, efficient backbones, and their combination with sparse attention or hierarchical attention techniques, it is possible to strike a balance between real-time performance and accuracy. The key trade-off is between reducing the computational load (for real-time performance) and maintaining model accuracy and consistency, which may require experimenting with different levels of simplification based on the specific use case. These advancements would facilitate deployment on real-world hardware and the development of a fully autonomous system capable of responding to real-time contingencies more efficiently. Additionally, the current datasets can be further enhanced by incorporating more balanced and diverse data. We anticipate that significant breakthroughs in this area will emerge in the coming years.

Author Contributions: Conceptualization, J.H. and C.F.; methodology, J.H.; software, J.H.; validation, Z.Z., X.L. and C.F.; formal analysis, J.H.; investigation, J.H.; resources, C.F.; data curation, J.H., Z.Z. and X.L.; writing—original draft preparation, J.H.; writing—review and editing, Z.Z., X.L. and C.F.; visualization, J.H.; supervision, Z.Z. and X.L.; project administration, Z.Z. and X.L.; funding acquisition, C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Aeronautical Science Foundation of China under grant 2024Z0710S5001.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to participants' privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tang, H.; Wang, Q.; Chen, H. Research on 3D human pose estimation using RGBD camera. In Proceedings of the 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 12–14 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 538–541.
2. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [\[CrossRef\]](#)
3. Rim, B.; Sung, N.J.; Ma, J.; Choi, Y.J.; Hong, M. Real-time human pose estimation using RGB-D images and deep learning. *J. Internet Comput. Serv.* **2020**, *21*, 113–121.
4. Pascual-Hernández, D.; de Frutos, N.O.; Mora-Jiménez, I.; Canas-Plaza, J.M. Efficient 3D human pose estimation from RGBD sensors. *Displays* **2022**, *74*, 102225. [\[CrossRef\]](#)
5. Serrano-Gotarredona, T.; Linares-Barranco, B. A 128×128 1.5% Contrast Sensitivity 0.9% FPN $3 \mu\text{s}$ Latency 4 mW Asynchronous Frame-Free Dynamic Vision Sensor Using Transimpedance Preamplifiers. *IEEE J. Solid-State Circuits* **2013**, *48*, 827–838. [\[CrossRef\]](#)
6. Brandli, C.; Berner, R.; Yang, M.; Liu, S.C.; Delbruck, T. A 240×180 130 db $3 \mu\text{s}$ latency global shutter spatiotemporal vision sensor. *IEEE J. Solid-State Circuits* **2014**, *49*, 2333–2341. [\[CrossRef\]](#)
7. Amir, A.; Taba, B.; Berg, D.; Melano, T.; McKinstry, J.; Di Nolfo, C.; Nayak, T.; Andreopoulos, A.; Garreau, G.; Mendoza, M.; et al. A low power, fully event-based gesture recognition system. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7243–7252.
8. Gallego, G.; Lund, J.E.; Mueggler, E.; Rebecq, H.; Delbruck, T.; Scaramuzza, D. Event-based, 6-DOF camera tracking from photometric depth maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2402–2412. [\[CrossRef\]](#)
9. Gehrig, D.; Rebecq, H.; Gallego, G.; Scaramuzza, D. Asynchronous, photometric feature tracking using events and frames. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 750–765.
10. Rebecq, H.; Gallego, G.; Mueggler, E.; Scaramuzza, D. EMVS: Event-based multi-view stereo—3D reconstruction with an event camera in real-time. *Int. J. Comput. Vis.* **2018**, *126*, 1394–1414. [\[CrossRef\]](#)
11. Jiang, J.; Li, J.; Zhang, B.; Deng, X.; Shi, B. Evhandpose: Event-based 3d hand pose estimation with sparse supervision. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 6416–6430. [\[CrossRef\]](#)
12. Ramakrishna, V.; Munoz, D.; Hebert, M.; Andrew Bagnell, J.; Sheikh, Y. Pose machines: Articulated pose estimation via inference machines. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part II 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 33–47.
13. Tompson, J.J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
14. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
15. Papandreou, G.; Zhu, T.; Chen, L.C.; Gidaris, S.; Tompson, J.; Murphy, K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–286.
16. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
17. Dosovitskiy, A. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
18. Peng, Y.; Yang, X.; Li, D.; Ma, Z.; Liu, Z.; Bai, X.; Mao, Z. Predicting flow status of a flexible rectifier using cognitive computing. *Expert Syst. Appl.* **2025**, *264*, 125878. [\[CrossRef\]](#)
19. Mao, Z.; Asai, Y.; Yamanoi, A.; Seki, Y.; Wiranata, A.; Minaminosono, A. Fluidic rolling robot using voltage-driven oscillating liquid. *Smart Mater. Struct.* **2022**, *31*, 105006. [\[CrossRef\]](#)
20. Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; Collomosse, J.P. Total capture: 3D human pose estimation fusing video and inertial sensors. In Proceedings of the BMVC, London, UK, 4–7 September 2017; Volume 2, pp. 1–13.
21. Zhang, Z.; Wang, C.; Qin, W.; Zeng, W. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2200–2209.
22. Zheng, Z.; Yu, T.; Li, H.; Guo, K.; Dai, Q.; Fang, L.; Liu, Y. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 384–400.
23. Roetenberg, D.; Luinge, H.; Slycke, P. Xsens MVN: Full 6DOF human motion tracking using miniature inertial sensors. *Xsens Motion Technol. BV Tech. Rep* **2009**, *1*, 1–7.
24. Von Marcard, T.; Rosenhahn, B.; Black, M.J.; Pons-Moll, G. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2017; Volume 36, pp. 349–360.

25. Von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G. Recovering accurate 3d human pose in the wild using imus and a moving camera. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 601–617.
26. Guзов, V.; Mir, A.; Sattler, T.; Pons-Moll, G. Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4318–4329.
27. Chen, H.; Leu, M.C.; Yin, Z. Real-time multi-modal human–robot collaboration using gestures and speech. *J. Manuf. Sci. Eng.* **2022**, *144*, 101007. [[CrossRef](#)]
28. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
29. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.
30. Tian, Z.; Chen, H.; Shen, C. Directpose: Direct end-to-end multi-person pose estimation. *arXiv* **2019**, arXiv:1911.07451.
31. Sun, X.; Shang, J.; Liang, S.; Wei, Y. Compositional human pose regression. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2602–2611.
32. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral human pose regression. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 529–545.
33. Nie, X.; Feng, J.; Zhang, J.; Yan, S. Single-stage multi-person pose machines. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6951–6960.
34. Wang, D. Stacked Dense-Hourglass Networks for Human Pose Estimation. Ph.D. Thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 2018.
35. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning feature pyramids for human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1281–1290.
36. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
37. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
38. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
39. Snower, M.; Kadav, A.; Lai, F.; Graf, H.P. 15 keypoints is all you need. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6738–6748.
40. He, Y.; Yan, R.; Fragkiadaki, K.; Yu, S.I. Epipolar transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7779–7788.
41. Lin, K.; Wang, L.; Liu, Z. End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1954–1963.
42. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3d human pose estimation with spatial and temporal transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11656–11665.
43. Stoffl, L.; Vidal, M.; Mathis, A. End-to-end trainable multi-instance pose estimation with transformers. *arXiv* **2021**, arXiv:2103.12115.
44. Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; Wang, Z. Tfpose: Direct human pose estimation with transformers. *arXiv* **2021**, arXiv:2103.15320.
45. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. Vitpose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38571–38584.
46. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11313–11322.
47. Baldwin, R.W.; Liu, R.; Almatrafi, M.; Asari, V.; Hirakawa, K. Time-ordered recent event (tore) volumes for event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2519–2532. [[CrossRef](#)] [[PubMed](#)]
48. Chen, J.; Shi, H.; Ye, Y.; Yang, K.; Sun, L.; Wang, K. Efficient human pose estimation via 3d event point cloud. In Proceedings of the 2022 International Conference on 3D Vision (3DV), Prague, Czechia, 12–15 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–10.
49. Calabrese, E.; Taverni, G.; Awai Easthope, C.; Skriabine, S.; Corradi, F.; Longinotti, L.; Eng, K.; Delbruck, T. DHP19: Dynamic vision sensor 3D human pose dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

51. Zou, S.; Guo, C.; Zuo, X.; Wang, S.; Wang, P.; Hu, X.; Chen, S.; Gong, M.; Cheng, L. Eventhpe: Event-based 3d human pose and shape estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10996–11005.
52. Goyal, G.; Di Pietro, F.; Carissimi, N.; Glover, A.; Bartolozzi, C. MoveEnet: Online high-frequency human pose estimation with an event camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4024–4033.
53. Duan, P.; Wang, Z.W.; Zhou, X.; Ma, Y.; Shi, B. EventZoom: Learning to denoise and super resolve neuromorphic events. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–25 June 2021; pp. 12824–12833.
54. Zou, Y.; Zheng, Y.; Takatani, T.; Fu, Y. Learning to reconstruct high speed and high dynamic range videos from events. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 21–25 June 2021; pp. 2024–2033.
55. Rehder, J.; Nikolic, J.; Schneider, T.; Hinzmann, T.; Siegwart, R. Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 4304–4311.
56. Rebecq, H.; Ranftl, R.; Koltun, V.; Scaramuzza, D. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1964–1980. [[CrossRef](#)] [[PubMed](#)]
57. Wang, J.; Tan, S.; Zhen, X.; Xu, S.; Zheng, F.; He, Z.; Shao, L. Deep 3D human pose estimation: A review. *Comput. Vis. Image Underst.* **2021**, *210*, 103225. [[CrossRef](#)]
58. Diederik, P.K. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
59. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph. (tog)* **2019**, *38*, 1–12. [[CrossRef](#)]
60. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. Hrformer: High-resolution transformer for dense prediction. *arXiv* **2021**, arXiv:2110.09408.
61. Liu, D.; Wang, T.; Sun, C. Voxel-based multi-scale transformer network for event stream processing. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 2112–2124. [[CrossRef](#)]
62. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.H.; Koltun, V. Point transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16259–16268.
63. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.