# Event-based Head Pose Estimation: Benchmark and Method

Jiahui Yuan[1,*], Hebei Li[1,*], Yansong Peng[1], Jin Wang[1], Yuheng Jiang[1], Yueyi Zhang[1,†], and Xiaoyan Sun[1,2]

[1] University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{yuanjiahui,lihebei,pengyansong,jin01wang,jiangyuheng}@mail.ustc.edu.cn
{zhyuey,sunxiaoyan}@ustc.edu.cn

**Abstract.** Head pose estimation (HPE) is crucial for various applications, including human-computer interaction, augmented reality, and driver monitoring. However, traditional RGB-based methods struggle in challenging conditions like sudden movement and extreme lighting. Event cameras, as a neuromorphic sensor, have the advantages of high temporal resolution and high dynamic range, offering a promising solution for HPE. However, the lack of paired event and head pose data hinders the full potential of event-based HPE. To address this, we introduce two large-scale, diverse event-based head pose datasets encompassing 282 sequences across different resolutions and scenarios. Furthermore, we propose the event-based HPE network, featuring two novel modules: the Event Spatial-Temporal Fusion (ESTF) module and the Event Motion Perceptual Attention (EMPA) module. The ESTF module effectively combines spatial and temporal information from the event streams, while the EMPA module captures crucial motion details across the scene using a large receptive field. We also propose a unified loss function to optimize the network using both angle and rotation matrix information. Extensive experiments demonstrate the superior performance of our network on both datasets, showcasing its effectiveness in handling HPE across various challenging scenarios. The datasets and code are available at https://github.com/Jiahui-Yuan-1/EVHPE.

**Keywords:** Head Pose Estimation · Event Camera

## 1 Introduction

Head pose estimation (HPE) has established itself as a crucial task in computer vision, tasked with determining the three-dimensional orientation of a person's head relative to a reference point, typically the camera. The ability to accurately estimate head pose unlocks a plethora of applications across various domains, such as human-computer interaction [39], driver monitoring systems [25] and augmented reality [19].

[*]Equal contribution [†]Corresponding author

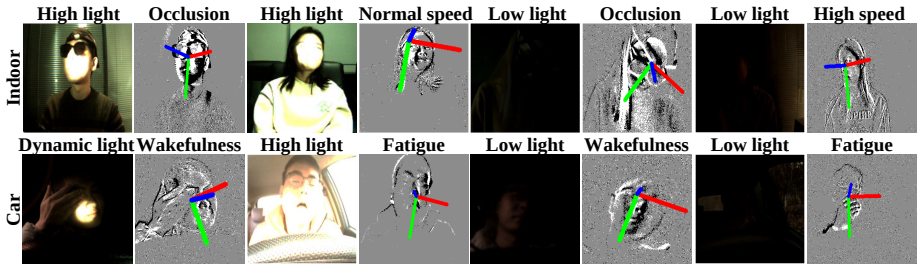


**Fig. 1:** Representative scenes from the Davis-HP dataset. The left column displays RGB frames, while the right column shows the event flow between RGB frames with annotations. The descriptions of the capturing setting are demonstrated above images.

Recent advancements in HPE have leveraged various strategies to improve accuracy. Some methods utilize facial landmark information to assist HPE [3,18], while others incorporate depth data to capture the three-dimensional nature of head movements, improving accuracy [6, 22–24]. Additionally, diverse learning strategies like stage-wise regression [44], multi-task learning [1, 41], and parametric orientation approaches [4,9,11,12,21] have further enhanced performance. However, despite these advancements, current HPE methods often require high-quality RGB input, which can still suffer limitations during rapid head movements and extreme lighting conditions.

Recently, event cameras [7], inspired by biological vision, are currently gaining increased attention from an expanding research community. Unlike RGB cameras that capture frames at fixed intervals, event cameras asynchronously detect changes in log intensity at the pixel level, capturing only the dynamic aspects of visual scenes. Therefore, event cameras have significant advantages, including high temporal resolution ($< 10\mu$s) and high dynamic range ($> 120$dB) [36]. These features give event cameras the potential to handle rapid head movements and extreme lighting scenarios in the HPE task.

However, existing HPE tasks still lack datasets that include event camera data. To fully leverage the advantages of event data in HPE, we have introduced two real-world head pose datasets named Prophesee-HP and Davis-HP. Unlike existing real-world datasets with low temporal resolution, our datasets offer different resolutions and diverse scenarios, along with detailed temporal annotations. These annotations are crucial for creating high temporal resolution datasets for event-based HPE. Fig. 1 shows several representative scenes from the Davis-HP dataset.

We present a novel event-based HPE (EV-HPE) network, which stands as a substantial advancement in the field. Our network incorporates an Event Spatial-Temporal Fusion (ESTF) module and Event Motion Perceptual Attention (EMPA) modules. The ESTF module is designed to enrich the spatial information across the temporal dimension. This module compresses the channels in different ways, aiming to capture and amplify the spatial details that are essential for accurate HPE. The EMPA module is applied to identify the crit-

ical area of head movements across diverse and complex scenarios. This module utilizes a large receptive field to guide the feature through attention map, thereby enhancing the performance of HPE. Additionally, we propose a unified loss that combines angle loss and geodesic loss to better optimize the network for event-based HPE. To validate the effectiveness, we evaluate our method on the Prophesee-HP and Davis-HP datasets. Experimental results show our significant effectiveness, establishing new benchmarks for the Prophesee-HP and Davis-HP datasets.

In brief, our contributions are summarized as follows:

- To the best of our knowledge, our work represents the first application of event cameras to the HPE task, offering potential advantages for handling rapid head movements and extreme lighting scenarios.
- To facilitate further research in HPE, we have constructed two large-scale event-based head pose datasets encompassing diverse scenarios.
- We propose the first event-based network for HPE, with two specially design modules and a unified loss.
- Experimental results validate the effectiveness of our proposed method on the proposed datasets.

## 2 Related Work

### 2.1 Head Pose Estimation

In the domain of the HPE task, early methods utilize extra information-utilized approaches to enhance the performance of HPE. Xin *et al.* [42] introduced a novel method that treats HPE as a graph regression problem, utilizing the graph neural network. However, the performance of HPE heavily depends on the accuracy of facial landmark detector. Wu *et al.* [41] introduced a SynergyNet that combines 3D Morphable Models (3DMM) and 3D facial landmarks to accurately predict complete facial geometry, thereby improving the accuracy of estimation. Besides, Fanelli *et al.* [6] introduced a system that estimate the position and head pose through additional depth information. Meyer *et al.* [23] proposed a robust method through registering 3D morphable models to depth images and iterative refining the registration over time.

Recently, researchers start to explore the effective learning strategies for HPE. Ruiz *et al.* [34] first achieved end-to-end prediction of three Euler angles through combination between classification and regression. Yang *et al.* [44] introduced a stage-wise regression method from a single image, employing feature aggregation with a fine-grained structure to enhance spatial feature grouping. Recently, Cao *et.al.* [4] proposed a vector-based head pose representation method that addresses the discontinuity issues of Euler angle annotations. To address the problem of ambiguous ground truth, Hempel *et al.* [11] introduced the 6D rotation matrix representation to restrict range of angle, thereby predicting satisfactory results. Zhang *et al.* [46] proposed a transformer-based approach utilizing orientation tokens for efficient and accurate head pose estimation. However, these

**Table 1:** Comparison with other real-world HPE datasets.

| Dataset | Resolution | #Scene | #Volunteer | #Seq. | #Frame | #Anno. | Source |
|---|---|---|---|---|---|---|---|
| AFLW2000 [50] | 450x450 | / | / | / | 2000 | 2000 | RGB |
| BIWI [6] | 640x480 | 1 | 20 | 24 | 15678 | 15678 | RGB+D |
| Davis-HP | 346x260 | 20 | 31 | 282 | 170636 | 683102 | RGB+E |
| Prophesee-HP | 1280x720 | 20 | 31 | 282 | / | 547653 | Event |

**Table 2:** The summary of the Prophesee-HP and Davis-HP datasets, including the number of slices (10ms) and event.

| | Prophesee-HP | | | Davis-HP | | |
|---|---|---|---|---|---|---|
| | Train | Test | Total | Train | Test | Total |
| #Slices (10ms) | 393655 | 153998 | 547653 | 492576 | 190526 | 683102 |
| #Event | 12.11G | 6.27G | 18.38G | 1.12G | 0.52G | 1.63G |

methods require high-quality RGB information, which could not handle the extremely lighting scenarios.

### 2.2 Event Camera and Applications

Different from conventional cameras, event cameras [13] asynchronously capture events when brightness changes exceed a certain threshold. Therefore, event cameras are particularly beneficial in scenarios requiring real-time and energy efficient systems, like robotics or edge devices, especially in environments with variable lighting. These cameras are applied in various applications, including object detection [26–28], surveillance [20], depth estimation [29, 32], optical flow detection [2, 48], HDR image reconstruction [31, 40], video frame interpolation [37, 45], video super resolution [16, 17] and Simultaneous Localization and Mapping (SLAM) [30, 38].

In event-based vision, event cameras are well-suited for pose estimation. Xu *et al.* [43] proposed a hybrid event camera method that generates both an asynchronous event stream and low-frequency grayscale images to capture high-frequency 3D volumetric poses. To capture local motion, Zou *et al.* [51] proposed a novel two-stage deep learning framework through optical flow and shape to estimate 3D human shapes. Goyal *et al.* [10] proposed a real-time system for high-frequency 2D human pose estimation using event cameras. Rudnev *et al.* [33] introduced a new approach for 3D hand pose estimation, which can be trained on synthetic event streams and then generalized to real-world data. However, event cameras have not been applied to HPE.

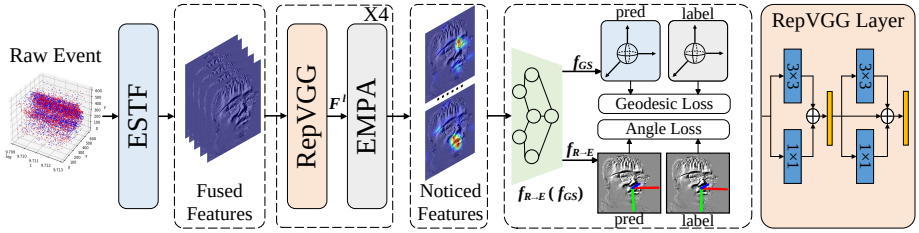


**Fig. 2:** Overview of the proposed method for event-based Head Pose Estimation.

## 3 Benchmark Dataset

To our knowledge, there are no datasets for event-based HPE. To explore the event-based HPE, we constructed event-based HPE data ourselves. Our data collection involved 31 volunteers, recording 15 participants in an indoor environment and 16 participants in a car. To capture diverse scenarios, we varied lighting conditions (strong, normal, low, extremely low, with added dynamic flickering light in the car), head movement speeds (from normal to high), and occlusions (eyeglasses, hats). Additionally, for the car recordings, we collected data under different driving states, including wakefulness and fatigue. This comprehensive approach ensures our event-based head pose datasets encompass a rich variety of real-world conditions.

To capture high-resolution head pose dynamics, volunteers wore an HWT906 gyroscope sensors on their heads, recording data at up to 1000 Hz. Before data collection, we calibrated these sensors with the volunteers' heads in a horizontal position. During recording, timestamps from the computer triggered synchronized data capture from all sensors. We used two event cameras (Prophesee EVK4 and DAVIS346) to capture data for each volunteer. This process resulted in the construction of two large-scale event-based HPE datasets: Prophesee-HP and Davis-HP. As highlighted in Tab. 1, our datasets offer a richer variety of scenarios and a larger scale compared to two commonly used real-world HPE benchmarks. Tab. 2 shows the total numbers of slice (10ms) and event in the Prohpesee-HP and Davis-HP datasets.

## 4 Method

### 4.1 Overview

Fig. 2 illustrates our novel event-based HPE network, EV-HPE. Its core components are: an Event Spatial-Temporal Fusion (ESTF) module, four lightweight RepVGG blocks (inspired by [5]), four Event Motion Perception Attention (EMPA) modules, a prediction head with multi-layer perceptrons (MLPs).

The event stream is initially transformed into a voxel grid representation [8] [49]. The ESTF module augments the temporal aspects of head motion in the data. This processed data is input into four feature extraction blocks, each

consisting of RepVGG layers and EMPA modules to capture essential motion information during head movements. After feature extraction, the network employs a perceptron and a matrix mapping layer to estimate the rotation matrix. This matrix is subsequently converted into Euler angles via an angle mapping technique. The predicted angles are then evaluated against ground truth labels to calculate both Geodesic and Angle Loss, which facilitate the training of the network.

### 4.2 Event Spatial-Temporal Fusion Module

The voxel grid representation of event information suffers from sparsity in the temporal dimension. This sparsity can hinder HPE tasks that require detailed understanding of spatial motion over time. To address this challenge, we enhance the spatial information within the original representation, allowing the network to learn more effective features for head pose estimation.

To achieve this, we propose an Event-based Spatio-Temporal Fusion (ESTF) module, as illustrated in Fig. 3(a). This module aims to enrich the representation of spatial information along the temporal dimension. The ESTF module first generates two separate representations by applying mean and max pooling operations along the temporal axis. Mathematically, this can be expressed as:

$$F_{Mean}\left(i,j\right)=\frac{1}{B}\sum_{k=1}^{B}F\left(i,j,k\right),F_{Max}\left(i,j\right)=Max_{k=1}^{B}F\left(i,j,k\right), \tag{1}$$

where $F_{Mean}$ and $F_{Max}$ represent the newly obtained representations, respectively. $F$ represents the voxel representation after processing the original event information. $F(i,j,k)$ denotes the value at position $(i,j)$ and the k-th temporal bin in the voxel grid. $B$ represents the number of bins in the voxel grid.

The mean operation smooths out the noise, preserving the differential information between pixels, while the max operation highlights detailed temporal features with relatively sharper noise. Subsequently, we extract horizontal and vertical edge contours from these two representations using the Adaptive Sobel Module, initially setting the kernel weights to Sobel X and Sobel Y operators, shown in Fig. 3(b). These weights can be dynamically adjusted with a minimal learning rate to better adapt to the event information stream. Finally, we concatenate both representations and their edge features with the original representation, thereby enhancing the spatial information of the original representation through a fusion block, as illustrated in Fig. 3(c). The fusion process can be expressed by the formula:

$$F_{fusion}=Fusion\left\{Concat\left[F,F_{\mathrm{w}},f_{s_x}\left(F_{\mathrm{w}}\right),f_{s_y}\left(F_{\mathrm{w}}\right)\right],F\right\}, \tag{2}$$

where $W$ is an operator, such as the Mean or Max operator, and $f_{s_x}$, $f_{s_y}$ represent the convolution functions with Sobel X and Sobel Y, respectively. $F_{fusion}$ denotes the fused feature, which is the output of the ESTF module.

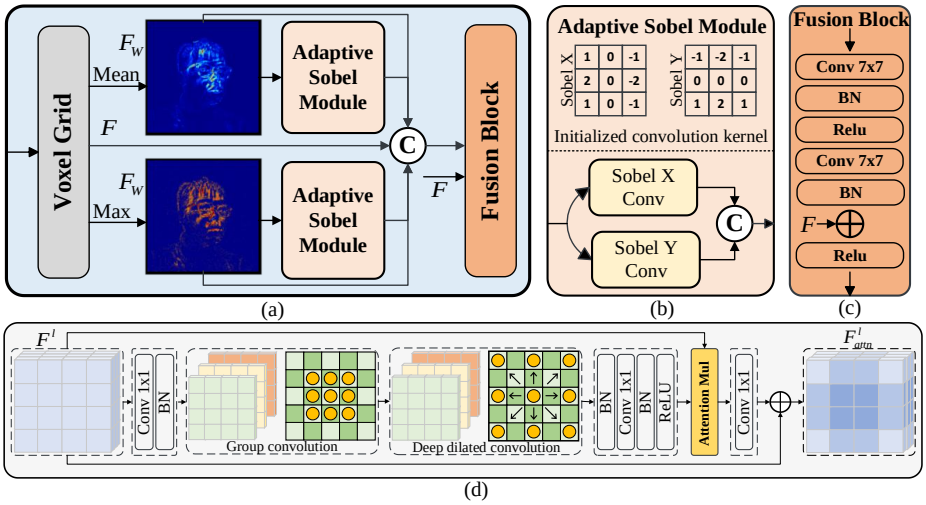


**Fig. 3:** (a) An overview of the Event Spatio-Temporal Fusion module (ESTF). (b) The components of the Adaptive Sobel module in ESTF. (c) The components of fusion block. (d) An overview of the Event Motion Perceptual Attention module (EMPA).

### 4.3 Event Motion Perceptual Attention module

To address the challenges of HPE in complex scenarios with rapid head movements, various lighting conditions and occlusion, we introduce an Event Motion Perception Attention (EMPA) module. This module enables the network to focus on learning the critical information of head movements across diverse and complex environments. Initially, as illustrated in Fig. 3(d), the event features generate a new feature map $F^l$ after passing through a 1x1 convolution layer followed by a GELU activation function. Then, $F^l$ is passed through a 1x1 convolution layer and batch normalization to further generalize the features. Next, we use a grouped local convolution layer. This layer lets the network learn features independently within each group, helping to represent a wider variety of features. This local convolution focuses on important positional information for pose estimation. Subsequently, we employ deep dilated convolution to capture more extensive contextual information. This approach improves accuracy and robustness in complex scenes by considering both local features and environmental context. Next, we normalize the feature map from the deep dilated convolution, apply a 1x1 convolution and batch normalization, followed by the ReLU activation function. The output serves as the attention weights $Attn$. We multiply the attention weights with the original features from the other pathway, pass them through a 1x1 convolution layer, and then apply a residual connection, enabling the network to learn crucial information about head movements across diverse scenarios. The process can be expressed by the following formula:

$$F^l_{attn} = f_{conv}\left(Attn \cdot F^l\right) + F^l, \tag{3}$$

where $F^l$ and $F^l_{attn}$ represent the input features and noticed features of the $l^{th}$ EMPA module, $f_{conv}$ stands for a 1x1 convolution layer, respectively.

### 4.4 Loss function

After obtaining the final feature map, our event data stream passes through an average pooling layer, followed by a multi-layer perceptron that regresses six parameters [11]. The prediction process of the rotation matrix can be mathematically represented as follows:

$$R_{\text{pred}} = f_{\text{GS}}\left(W_2\left(\text{LeakyReLU}\left(W_1 X + b_1\right)\right) + b_2\right), \tag{4}$$

where $R_{pred}$ denotes the predicted rotation matrix, $f_{GS}(\cdot)$ represents the mapping of two column matrices into the rotation matrix, and $W_1$, $W_2$, $b_1$, $b_2$ represent the parameters in the perceptron's linear layers. LeakyReLU denotes the activation function.

In this paper, we employ the geodesic loss [11] as the primary loss function, which forms a rotation matrix using the six regressed parameters. The discrepancy between matrices is then measured by the geodesic distance between the predicted and true matrices. However, due to the diversity of evaluation metrics, we introduce an angle loss to better regress the Euler angle that captures the difference between the angles derived from the matrix transformation and the true angles, improving the prediction of the model. The geodesic loss $L_R$ and angle loss $L_E$ can be mathematically expressed as follows:

$$L_R = \cos^{-1}\left(\frac{tr\left(R_{pred}R_{gt}^T\right) - 1}{2}\right), \tag{5}$$

$$L_E = \sum_{i=1}^{3} \left| f_{R\to E}\left(R_{pred}\right)(i) - E_{gt}\left(i\right) \right|, \tag{6}$$

where $tr(\cdot)$ stands for the trace of a matrix. $R_{\text{pred}}$ and $R_{\text{gt}}$ represent the predicted rotation matrix and the ground truth rotation matrix, respectively. $f_{R\to E}(\cdot)$ represents the function that converts the rotation matrix to Euler angles, and $i$ indicates the corresponding Euler angle index. Therefore, the total loss function of the network is depicted as follow:

$$L_{total} = L_R + \alpha L_E, \tag{7}$$

where $\alpha$ represents the weight coefficient for the angle loss.

## 5 Experimental Results

### 5.1 Implementation Details

For model optimization, we employed the Adam optimizer with a learning rate of 5e-5 and a weight decay of 0.5. We used a batch size of 40 and trained for 20

**Table 3:** On the Prophesee-HP dataset, the input data format consists of 10ms event time slices. A quantitative comparison between existing HPE methods and our approach has been conducted, with the best results highlighted in bold.

| Method | Input | Euler angle errors | | | | Vector errors | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (10ms) | Pitch | Yaw | Roll | MAE | Left | Down | Front | MAEV |
| HopeNet [34] | Event | 9.20 | 9.30 | 5.95 | 8.15 | 12.71 | 11.99 | 14.41 | 13.04 |
| FSA-Net [44] | Event | 10.39 | 11.07 | 6.34 | 9.27 | 14.36 | 13.31 | 16.65 | 14.77 |
| WHENet [47] | Event | 8.72 | 10.03 | 5.98 | 8.24 | 13.11 | 11.54 | 14.59 | 13.08 |
| 6DRepNet [11] | Event | 7.00 | 8.44 | 5.11 | 6.85 | 11.12 | 9.42 | 11.95 | 10.83 |
| TokenHPE [46] | Event | 11.13 | 12.06 | 7.08 | 10.09 | 15.72 | 14.24 | 17.97 | 15.98 |
| Ours | Event | **6.00** | **7.47** | **4.55** | **6.01** | **9.89** | **8.20** | **10.49** | **9.53** |

**Table 4:** On the Davis-HP dataset, the input data format consists of 10ms event time slices. A quantitative comparison between existing HPE methods and our approach has been conducted, with the best results highlighted in bold.

| Method | Input | Euler angle errors | | | | Vector errors | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (10ms) | Pitch | Yaw | Roll | MAE | Left | Down | Front | MAEV |
| HopeNet [34] | Event | 8.98 | 9.98 | 5.95 | 8.30 | 13.08 | 11.71 | 14.89 | 13.23 |
| FSA-Net [44] | Event | 9.56 | 9.81 | 6.23 | 8.53 | 13.02 | 12.31 | 14.97 | 13.43 |
| WHENet [47] | Event | 8.50 | 9.14 | 5.81 | 7.82 | 12.23 | 11.13 | 13.70 | 12.35 |
| 6DRepNet [11] | Event | 8.30 | 8.62 | 4.91 | 7.28 | 10.97 | 10.38 | 13.15 | 11.50 |
| TokenHPE [46] | Event | 11.05 | 11.02 | 6.11 | 9.39 | 14.08 | 13.53 | 17.09 | 14.90 |
| Ours | Event | **7.50** | **7.57** | **4.41** | **6.49** | **9.75** | **9.32** | **11.65** | **10.24** |

epochs. The coefficient $\alpha$ in the total loss function was set to 0.001. To ensure a fair comparison with other methods, we excluded samples exceeding 99° or below -99° from both datasets due to the 198-bin limitation in HopeNet [34]. The training set comprised sequences from 22 volunteers, while the testing set included data from 9 volunteers.

For experiments with the Prophesee-HP dataset, we leveraged the recorded facial bounding box information to crop the input data to the resolution of 600×600. In contrast, for the Davis-HP dataset, all inputs were of size 346×260. Finally, to maintain a consistent and fair experimental platform, all training was conducted using 8 NVIDIA A800 GPU cards.

### 5.2 Experimental Results

**Quantitative Results.** Tab. 3 provides a detailed quantitative analysis of our network in comparision with other methods on the Prophesee-HP dataset. The comparison results show that our method surpasses current HPE methods for the same 10ms slice of event data stream, reducing average Euler angle errors by **0.84** and average vector errors by **1.30**. Particularly, our approach significantly

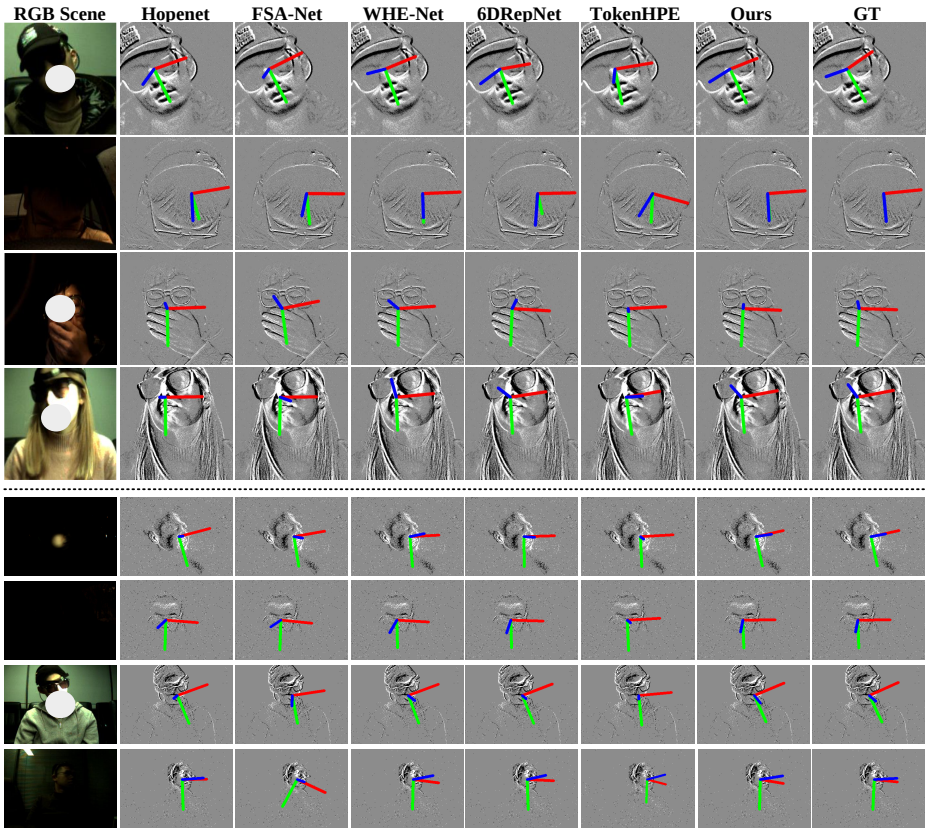


**Fig. 4:** Qualitative HPE results for four scene sequences using 10ms event slices. The top row showcases sequences from the Prophesee-HP dataset, while the bottom row presents sequences from the Davis-HP dataset. The leftmost column displays schematic diagrams depicting the captured volunteer scenes. The middle columns compare the performance of existing HPE methods with our proposed approach. Finally, the rightmost column presents the ground-truth angle labels.

enhances pitch and yaw angle predictions compared to the best-performing methods. This improvement is credited to the broad variability in pitch and yaw angles. Our ESTF module effectively integrates spatial information over time, while the EMPA module efficiently captures motion information of critical parts.

Tab. 4 shows the quantitative comparison between our method and other methods on the Davis-HP dataset. Our EV-HPE is superior to other methodologies when dealing with 10ms slices of data. The comparative analysis clearly illustrates that our approach outperforms the existing HPE techniques, with a notable decrease in the average Euler angle errors and average vector errors by **0.79** and **1.26**, respectively. Importantly, our method maintains a consistent edge in reducing angle errors across datasets of different resolutions. This advan-

**Table 5:** Quantitative comparison between existing HPE methods and our method on the Davis-HP dataset. To keep with the RGB frame rate, we choose 40ms slice for event data. The best results are highlighted in bold.

| Method | Input | Euler angles errors | | | | Vector errors | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (40ms) | Pitch | Yaw | Roll | MAE | Left | Down | Front | MAEV |
| HopeNet [34] | Event | 9.09 | 9.71 | 6.59 | 8.46 | 13.25 | 12.06 | 14.58 | 13.30 |
| | RGB | 14.11 | 13.95 | 7.67 | 11.91 | 17.32 | 17.16 | 21.92 | 18.80 |
| FSA-Net [44] | Event | 10.76 | 9.85 | 6.01 | 8.87 | 12.69 | 12.49 | 14.63 | 13.27 |
| | RGB | 13.32 | 12.94 | 6.61 | 10.96 | 16.07 | 15.92 | 20.25 | 17.41 |
| WHENet [47] | Event | 8.73 | 9.49 | 5.98 | 8.07 | 12.53 | 11.41 | 14.18 | 12.71 |
| | RGB | 12.05 | 13.60 | 6.66 | 10.77 | 16.37 | 15.14 | 20.20 | 17.24 |
| 6DRepNet [11] | Event | 8.29 | 8.59 | 5.09 | 7.32 | 11.04 | 10.49 | 13.13 | 11.55 |
| | RGB | 10.91 | 11.78 | 6.23 | 9.64 | 14.57 | 13.52 | 17.65 | 15.25 |
| TokenHPE [46] | Event | 11.23 | 10.30 | 5.90 | 9.14 | 13.43 | 13.87 | 16.73 | 14.68 |
| | RGB | 13.80 | 12.18 | 6.53 | 10.84 | 15.16 | 16.35 | 20.42 | 17.31 |
| Ours | Event | **6.97** | **7.62** | **4.80** | **6.46** | **9.93** | **9.05** | **11.34** | **10.11** |

tage is largely due to the proficiency of our EMPA module in capturing essential motion details across various scales.

Within the Davis-HP dataset, we specifically contrasted the event data stream between RGB frames with the test results from different methods applied to RGB frames. Tab. 5 provides a detailed quantitative analysis of our network's performance on inter-frame event data compared to other methods' performances on both inter-frame event data and RGB frame configurations. As observed from the table, existing HPE methods exhibit low accuracy due to the complexity and challenges presented by the scene. However, event data streams are able to compensate for the insufficiency of RGB information, particularly in conditions with varied luminance and velocity, thereby enhancing the precision of existing HPE methods in complex settings. Moreover, the comparative results unequivocally demonstrate that our method remains superior to current HPE techniques in handling the same inter-frame event data streams, reducing average Euler angle errors and average vector errors by **0.86** and **1.44**, respectively. This confirms that our network is capable of achieving optimal precision across different temporal segments of event streams

**Qualitative Results** Fig. 4 showcases the qualitative comparision of our proposed method against other methods on the Prophesee-HP dataset. We can observe that when volunteers wear occlusions and exhibit significant roll inclinations along with certain yaw angles, methods like 6DRepNet [11], HopeNet [34], and FSA-Net [44] fail to accurately estimate the yaw component, whereas WHENet [47] and TokenHPE [46] struggle with roll and pitch estimations. In scenarios where volunteers are in poorly lit environments, exhibiting actions akin to drowsy driving (such as lowering the head or hand obstructions), other methods

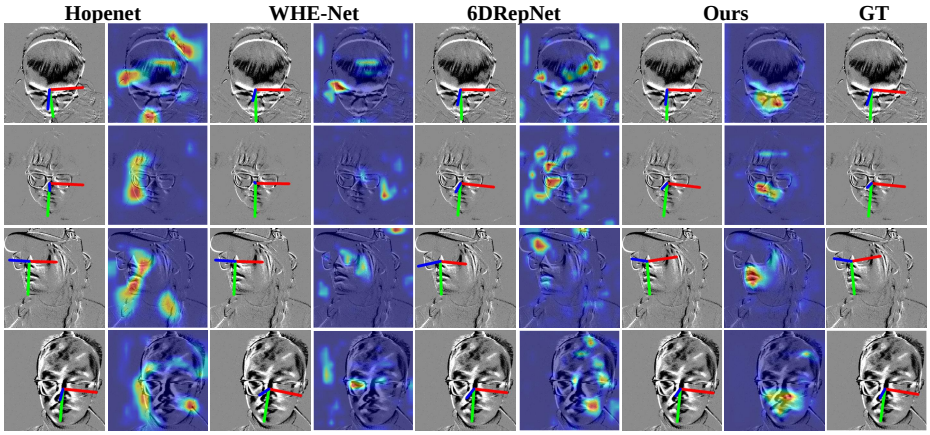


**Fig. 5:** Visualization of attention maps and HPE results generated by our method and other methods on the Prophesee-HP dataset.

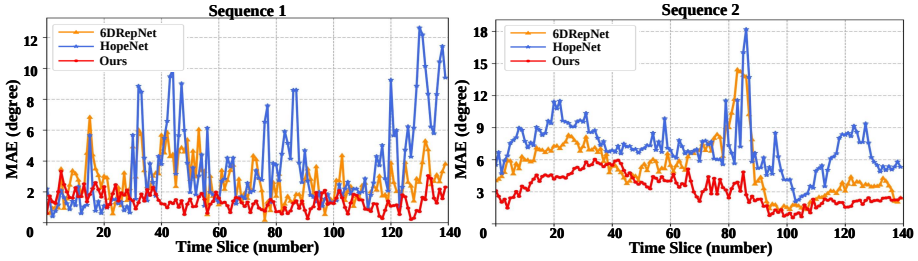


**Fig. 6:** Comparative stability analysis of the HPE results across two sequences on the Prophesee-HP dataset.

fall short in predicting the pitch angle. In contrast, our method effectively estimates both minor and significant pitch movements, attributed to the enhanced spatial motion information by our ESTF module. When faces are exposed to strong lighting, HopeNet [34], FSA-Net [44], WHENet [47], and 6DRepNet [11] inaccurately estimate the roll angle (indicated in red), and TokenHPE [46] has considerable deviation in yaw angle estimation. Our method closely aligns with the true labels for all three angles.

Fig. 4 illustrates qualitative results between our method and other approaches on the Davis-HP dataset. We can observe that our method in low-light conditions with dynamic lighting, facial halos predict accurate estimations, leading to significant errors in other methods. However, our approach provides more accurate estimations of yaw and subtle pitch movements. Under extremely low lighting, where event information is sparse, actions resembling fatigue lead to considerable yaw estimation errors in HopeNet [34], FSA-Net [44], WHE-Net, and TokenHPE [46], while 6DRepNet's [11] pitch estimation is less accurate compared to our method. During rapid head movements, which include both noise

and motion information, our method offers a more precise estimation of yaw angles with significant changes (highlighted in blue). In bright light conditions with some obstructions, our method aligns closest to the true labels despite the presence of interfering information.

### 5.3 Attention Map Visualization

To understand how the EMPA module focuses on critical pose information, we visualized the attention maps using GradCAM [35] during head pose prediction. We compared our method's attention patterns with leading HPE approaches like 6DRepNet [11], WHENet [47], and HopeNet [34]. Fig. 5 showcases two scenarios: indoor and in-car. We observe that our method effectively locates key positional information even under sparse event data conditions, unlike other methods that exhibit more dispersed attention in similar scenarios. In the third scenario, where hats and sunglasses obscure facial features, our method focuses on the nose and mouth, crucial for predicting three euler angles. During rapid head movements (the first and last scenarios), our method prioritizes attention on the nose and mouth, as these features are critical for pose estimation according to prior works [14, 15].

### 5.4 Stability Analysis

To delve deeper into the stability of our method's predictions over continuous time, we compared it with the two best-performing methods tested on the Prophesee-HP dataset. We evaluated their performance using Mean Absolute Error (MAE) across two continuous sequences, each containing 140 frames. A lower MAE value signifies a smaller deviation from the ground truth value. As illustrated in Fig. 6, our method consistently achieves lower MAE values across both sequences, indicating smaller overall deviations from the true head pose. While some fluctuations are present in all methods, our approach exhibits smoother and more stable performance compared to the other two existing HPE methods. This improved stability translates to a more reliable prediction of head pose over time, making it well-suited for real-world applications that require continuous and accurate tracking.

### 5.5 Ablation Study

**The effectiveness of ESTF, EMPA and Angle loss.** To evaluate the contributions of the ESTF and EMPA modules to our event-based HPE network, we conducted ablation experiments on the Prophesee-HP dataset. As shown in Tab. 6, removing the ESTF module significantly impacts performance. The lack of ESTF leads to sparse spatial information in the temporal dimension, hindering the capture of overall spatial context. Consequently, the EMPA module struggles to identify key spatial features, resulting in an MAE increase of approximately **0.51** and an MAEV increase of approximately **0.78**. Similarly,

**Table 6:** Ablation experimental results of assessing the effectiveness of different modules and losses on the Prophesee-HP dataset.

| Method | MAE | MAEV |
|---|---|---|
| w/o ESTF | 6.52 | 10.31 |
| w/o EMPA | 6.55 | 10.32 |
| w/o Angle Loss | 6.14 | 9.71 |
| ours | **6.01** | **9.53** |

**Table 7:** Ablation experimental results of evaluating the impact of the number of EMPA modules on the Prophesee-HP dataset.

| #EMPA Blk. | MAE | MAEV |
|---|---|---|
| 0 | 6.55 | 10.32 |
| 1 | 6.43 | 10.14 |
| 2 | 6.29 | 9.96 |
| 3 | 6.15 | 9.73 |
| 4 | **6.01** | **9.53** |

removing the EMPA module diminishes the network's ability to capture crucial motion information. This translates to an MAE increase of approximately **0.54** and an MAEV increase of approximately **0.79**. Finally, removing the Angle loss from the network leads to a slight decrease in overall regression accuracy, with the average Euler Angle error increasing by approximately **0.13** and the average vector error increasing by approximately **0.18**. These results highlight the critical roles of both ESTF and EMPA in extracting spatial and motion features from event streams, ultimately leading to more accurate head pose estimation.

**The impact of EMPA block number.** To explore the influence of the number of EMPA blocks on our network's performance, we conducted ablation experiments by incorporating varying numbers of stacked EMPA blocks. As shown in Tab. 7, adding a single EMPA block yields a notable improvement in performance. This suggests that the EMPA module effectively captures inter-layer flow information. Furthermore, stacking additional blocks progressively enhances this capability, leading to a gradual increase in performance.

## 6 Conclusion

This paper introduces the use of event cameras for HPE, offering a promising new approach in the field. We introduce the first event-based HPE network, accompanied by two large-scale, real-world event-based HPE datasets (Prophesee-HP and Davis-HP). These datasets capture challenging scenarios with high exposure, low light, rapid movements, and occlusions, pushing the boundaries of conventional HPE methods. To address these complexities, we design two novel modules, the Event Spatio-Temporal Fusion module and the Event Motion Perception Attention module, that demonstrably enhance HPE accuracy across diverse scenes. Rigorous testing on our established datasets showcases the superiority of our approach. This work represents a promising advancement in applying event cameras to HPE, providing more possibilities for their application.

## Acknowledgements

This work was in part supported by the National Natural Science Foundation of China (NSFC) under grants 62032006 and 62021001.

## References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7617–7627 (2021) 2
2. Benosman, R., Clercq, C., Lagorce, X., Ieng, S.H., Bartolozzi, C.: Event-based visual flow. IEEE transactions on neural networks and learning systems **25**(2), 407–417 (2013) 4
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE international conference on computer vision. pp. 1021–1030 (2017) 2
4. Cao, Z., Chu, Z., Liu, D., Chen, Y.: A vector-based representation to enhance head pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on applications of computer vision. pp. 1188–1197 (2021) 2, 3
5. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13733–13742 (2021) 5
6. Fanelli, G., Weise, T., Gall, J., Van Gool, L.: Real time head pose estimation from consumer depth cameras. In: Joint pattern recognition symposium. pp. 101–110. Springer (2011) 2, 3, 4
7. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020) 2
8. Gehrig, D., Loquercio, A., Derpanis, K.G., Scaramuzza, D.: End-to-end learning of representations for asynchronous event-based data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5633–5643 (2019) 5
9. Geng, X., Xia, Y.: Head pose estimation based on multivariate label distribution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1837–1842 (2014) 2
10. 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. pp. 4023–4032 (2023) 4
11. Hempel, T., Abdelrahman, A.A., Al-Hamadi, A.: 6d rotation representation for unconstrained head pose estimation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2496–2500. IEEE (2022) 2, 3, 8, 9, 11, 12, 13
12. Hsu, H.W., Wu, T.Y., Wan, S., Wong, W.H., Lee, C.Y.: Quatnet: Quaternion-based head pose estimation with multiregression loss. IEEE Transactions on Multimedia **21**(4), 1035–1046 (2018) 2
13. Indiveri, G., Douglas, R.: Neuromorphic vision sensors. Science **288**(5469), 1189–1190 (2000) 4

14. Iskra, A., Tomc, H.G.: Eye-tracking analysis of face observing and face recognition. Journal of Graphic Engineering and Design **7**(1), 5–11 (2016) 13
15. Jiang, M., Francis, S.M., Srishyla, D., Conelea, C., Zhao, Q., Jacob, S.: Classifying individuals with asd through facial emotion recognition and eye-tracking. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 6063–6068. IEEE (2019) 13
16. Jing, Y., Yang, Y., Wang, X., Song, M., Tao, D.: Turning frequency to resolution: Video super-resolution via event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7772–7781 (2021) 4
17. Kai, D., Zhang, Y., Sun, X.: Video super-resolution via event-driven temporal alignment. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 2950–2954. IEEE (2023) 4
18. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1867–1874 (2014) 2
19. Kumar, A., Alavi, A., Chellappa, R.: Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition (Feb 2017) 1
20. Litzenberger, M., Kohn, B., Belbachir, A.N., Donath, N., Gritsch, G., Garn, H., Posch, C., Schraml, S.: Estimation of vehicle speed based on asynchronous data from a silicon retina optical sensor. In: 2006 IEEE intelligent transportation systems conference. pp. 653–658. IEEE (2006) 4
21. Liu, H., Fang, S., Zhang, Z., Li, D., Lin, K., Wang, J.: Mfdnet: Collaborative poses perception and matrix fisher distribution for head pose estimation. IEEE Transactions on Multimedia **24**, 2449–2460 (2021) 2
22. Martin, M., Van De Camp, F., Stiefelhagen, R.: Real time head model creation and head pose estimation on consumer depth cameras. In: 2014 2nd International Conference on 3D Vision. vol. 1, pp. 641–648. IEEE (2014) 2
23. Meyer, G.P., Gupta, S., Frosio, I., Reddy, D., Kautz, J.: Robust model-based 3d head pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 3649–3657 (2015) 2, 3
24. Mukherjee, S.S., Robertson, N.M.: Deep head pose: Gaze-direction estimation in multimodal video. IEEE Transactions on Multimedia **17**(11), 2094–2107 (2015) 2
25. Murphy-Chutorian, E., Doshi, A., Trivedi, M.M.: Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In: 2007 IEEE intelligent transportation systems conference. pp. 709–714. IEEE (2007) 1
26. Peng, Y., Zhang, Y., Xiao, P., Sun, X., Wu, F.: Better and faster: Adaptive event conversion for event-based object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2056–2064 (2023) 4
27. Peng, Y., Zhang, Y., Xiong, Z., Sun, X., Wu, F.: Get: group event transformer for event-based vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6038–6048 (2023) 4
28. Perot, E., De Tournemire, P., Nitti, D., Masci, J., Sironi, A.: Learning to detect objects with a 1 megapixel event camera. Advances in Neural Information Processing Systems **33**, 16639–16652 (2020) 4
29. Rebecq, H., Gallego, G., Mueggler, E., Scaramuzza, D.: Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. International Journal of Computer Vision **126**(12), 1394–1414 (2018) 4

30. Rebecq, H., Horstschäfer, T., Gallego, G., Scaramuzza, D.: Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. IEEE Robotics and Automation Letters **2**(2), 593–600 (2016) 4
31. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE transactions on pattern analysis and machine intelligence **43**(6), 1964–1980 (2019) 4
32. Rogister, P., Benosman, R., Ieng, S.H., Lichtsteiner, P., Delbruck, T.: Asynchronous event-based binocular stereo matching. IEEE Transactions on Neural Networks and Learning Systems **23**(2), 347–353 (2011) 4
33. Rudnev, V., Golyanik, V., Wang, J., Seidel, H.P., Mueller, F., Elgharib, M., Theobalt, C.: Eventhands: Real-time neural 3d hand pose estimation from an event stream. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12385–12395 (2021) 4
34. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 2074–2083 (2018) 3, 9, 11, 12, 13
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017) 13
36. Serrano-Gotarredona, T., Linares-Barranco, B.: A 128 *times* 128 1.5% contrast sensitivity 0.9% fpn 3 $\mu$s latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. IEEE Journal of Solid-State Circuits **48**(3), 827–838 (2013) 2
37. Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D.: Time lens: Event-based video frame interpolation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16155–16164 (2021) 4
38. Vidal, A.R., Rebecq, H., Horstschaefer, T., Scaramuzza, D.: Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. IEEE Robotics and Automation Letters **3**(2), 994–1001 (2018) 4
39. Wang, Y., Liang, W., Shen, J., Jia, Y., Yu, L.F.: A deep coarse-to-fine network for head pose estimation from synthetic data. Pattern Recognition **94**, 196–206 (2019) 1
40. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2563–2572 (2021) 4
41. Wu, C.Y., Xu, Q., Neumann, U.: Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In: 2021 International Conference on 3D Vision (3DV). pp. 453–463. IEEE (2021) 2, 3
42. Xin, M., Mo, S., Lin, Y.: Eva-gcn: Head pose estimation based on graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 1462–1471 (2021) 3
43. Xu, L., Xu, W., Golyanik, V., Habermann, M., Fang, L., Theobalt, C.: 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. pp. 4968–4978 (2020) 4
44. Yang, T.Y., Chen, Y.T., Lin, Y.Y., Chuang, Y.Y.: Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1087–1096 (2019) 2, 3, 9, 11, 12

45. Yu, Z., Zhang, Y., Liu, D., Zou, D., Chen, X., Liu, Y., Ren, J.S.: Training weakly supervised video frame interpolation with events. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14589–14598 (2021) 4
46. Zhang, C., Liu, H., Deng, Y., Xie, B., Li, Y.: Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8897–8906 (2023) 3, 9, 11, 12
47. Zhou, Y., Gregson, J.: Whenet: Real-time fine-grained estimation for wide range head pose. arXiv preprint arXiv:2005.10353 (2020) 9, 11, 12, 13
48. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. arXiv preprint arXiv:1802.06898 (2018) 4
49. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019) 5
50. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016) 4
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. pp. 10996–11005 (2021) 4