IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

ExFMan: Rendering 3D Dynamic Humans with Hybrid Monocular Blurry Frames and Events

Kanghao Chen, Zeyu Wang, and Lin Wang

***Abstract*—Recent advances in neural rendering have enabled the 3D reconstruction of dynamic humans from monocular videos, with applications in robotics. However, it is still challenging to reconstruct clear humans from in-the-wild video encountering motion blur, causing shape and appearance inconsistencies, especially in blurry regions like hands and legs. In this paper, we propose ExFMan, the *first* neural rendering framework that unveils the possibility of rendering high-quality humans in rapid motion with a hybrid frame-based RGB and bio-inspired event camera. The "out-of-the-box" insight is to leverage the high temporal information of event data in a complementary manner and adaptively reweight the effect of losses for both RGB frames and events in the local regions, according to the velocity of the rendered human. This significantly mitigates the *inconsistency* associated with motion blur in the RGB frames. Specifically, we first formulate a velocity field of the 3D body in the canonical space and render it to image space to identify the body parts with motion blur. We then propose two novel losses, *i.e.*, velocity-aware photometric loss and velocity-relative event loss, to optimize the neural human for both modalities under the guidance of the estimated velocity. In addition, we incorporate novel pose regularization and alpha losses to facilitate continuous pose and clear boundary. Extensive experiments on synthetic and real-world datasets demonstrate that ExFMan can reconstruct sharper and higher quality humans over the compared baselines and the state-of-the-art methods for diverse blurry subjects.**

***Index Terms*—Sensor Fusion, Modeling and Simulating Humans, Visual Learning.**

I. Introduction

DIGITAL modeling of humans holds great potential for applications in virtual reality and robotics, where human models serve as intuitive interfaces. Rendering humans from monocular videos has gained attention due to the widespread availability of the data and their use in complex scenes [1]. Recent works [2], [3], [4] have focused on learning dynamic humans by deforming neural radiance fields (NeRF) [5] and 3D Gaussian splatting (3DGS) [6]. While these methods excel in controlled environments with ideal conditions, they struggle in in-the-wild scenarios with free-form human actions and varying lighting. Motion blur, common in such settings, degrades the quality of affected regions, causing inconsistencies during body optimization (see Fig. 1(top)).

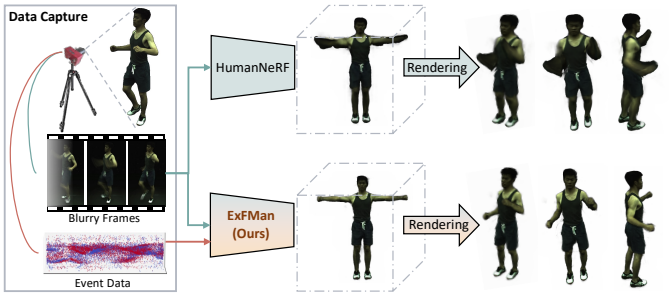

Fig. 1. Dynamic human motion often induces severe blur in videos, complicating the optimization process. **Left**: Data capture system. **Top row**: The baseline method optimizes only on RGB, ignoring motion blur, leading to blurry rendering. **Bottom row**: Our method integrates both RGB and event data in a novel pipeline, producing high-quality rendering.

Therefore, it is imperative to reconstruct 3D dynamic humans from the in-the-wild monocular blurry video; however, this task is non-trivial for a key reason – **inconsistency of human shape and appearance**. Existing methods [1], [7] optimize the dynamic human in a single static canonical space and deform it to fit multiple frames, which essentially relies on the underlying assumption of consistency of overall frames. Although some methods [8], [9], [10] reconstruct NeRF from blurry frames, they mainly focus on the general scene and only consider the blur from camera shift or defocus. This renders it difficult to apply them to reconstruct dynamic humans directly. To our knowledge, ***no prior work has directly addressed the challenge of rendering dynamic humans from the monocular blurry video***. A naive solution is to combine the video deblurring methods, *e.g.*, MPR [11] with the human rendering methods, *e.g.*, HumanNeRF [1] based on the deblurred frames in a two-stage manner. However, such a solution heavily relies on deblurring methods, which suffer from limited generalization capability across different scenes and often fail with complex motion. Consequently, it leads to error-prone appearance and shape in the rendered humans, as demonstrated in Fig. 4.

Event cameras are bio-inspired sensors that capture per-pixel intensity changes asynchronously. Recently, they are becoming popular for their distinct advantages, *e.g.*, high temporal resolution and high dynamic range (HDR). Accordingly, event cameras have been applied to tasks like video deblurring [12], [13], [14] and low-light enhancement [15],

Received 22 January 2025; accepted 11 June 2025. Date of publication 26 June 2025; date of current version 9 July 2025. This article was recommended for publication by Associate Editor B. TAMADAZTE and Editor P. Vasseur upon evaluation of the reviewers' comments. This work was supported in part by the Science Foundation of China (NSF) under Grant 62206069 (affiliated with Guangzhou HKUST Fok Ying Tung Research Institute), in part by the MOE AcRF Tier 1 SSHR-TG Incubator Grant FY24, under Grant RSTG7/24, and in part by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007. (Corresponding author: Lin Wang.)

Kanghao Chen and Zeyu Wang are with the The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China (e-mail: kchen879@connect.hkust-gz.edu.cn; zeyuwang@ust.hk).

Lin Wang is with Nanyang Technological University, Singapore 639798 (email: linwang@ntu.edu.sg).

The codes and datasets are publicly available in https://github.com/KHao123/ExFMan.

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3583602, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3583602

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

[16], [17]. They have also been leveraged to capture 3D human motion [18], [19] or estimate the human pose and shape [20], [21], by leveraging the advantages of event data.

In this paper, we explore a promising direction of rendering 3D dynamic humans with hybrid monocular blurry RGB frames and events. Intuitively, we introduce **ExFMan**, the ***first*** neural rendering framework that can render dynamic humans with high-quality appearance under diverse blur conditions. Our method can achieve high-quality novel view synthesis, as depicted in Fig. 1 (bottom). The key insight is to leverage the high temporal information of events in a complementary manner and adaptively reweight the effect of losses for both RGB frames and events in the local regions, according to the velocity of the rendered human. This significantly mitigates the ***inconsistency*** associated with motion blur in the RGB frames. To achieve this, we first formulate a velocity field of a 3D body within the canonical space, which is then deformed and rendered to the image space to identify the body regions with motion blur (Sec. III-B). Based on the estimated velocity, we propose two velocity-based rendering losses to facilitate the joint optimization for both modalities (Sec. III-C). Specifically, a velocity-aware photometric loss is designed by representing the rendered color through a Gaussian distribution, with the velocity score representing the variance. This approach directly diminishes the impact of the blurry regions, thus preserving the consistency of the human body across video frames. The body regions (*e.g.*, hands, legs) that always exhibit motion blur are not adequately optimized by the proposed photometric loss. Therefore, we apply additional supervision by designing a velocity-relative event loss for these under-constrained regions. Regarding the velocity as prediction uncertainty, the novel event loss optimizes the high-velocity regions by anchoring them to regions with lower velocity. Additionally, we incorporate the pose regularization and velocity-based alpha loss to facilitate the continuous pose and clear boundary (Sec. III-D).

Our contributions can be summarized as follows: **1)** We introduce the **first** method for rendering dynamic humans in rapid motion using hybrid monocular blurry RGB frames and sparse events **2)** We propose a novel framework with an event-oriented, blur-aware velocity field and two velocity-aware rendering losses. **3)** We unveil ExFMan with significant superiority on existing datasets.

II. Related Work

Neural Rendering for Human Reconstruction. Following the advent of Neural Radiance Fields (NeRF) [5], significant progress has been made in high-fidelity rendering for static scenes [23] and moving subjects [24], [25]. Recent advances have applied NeRF to human reconstruction in diverse scenarios, including monocular video [1], [26], rendering efficiency [27], [28] and animation [7]. Our work follows HumanNeRF [1], known for its excellent rendering from monocular video. HumanNeRF uses a static T-pose as the canonical model and applies a deformation mapping [29] to adapt it to the observation space of each video frame (details in Sec. III-A). Recently, the 3D Gaussian splitting (3DGS) [6] has gained attention for balancing real-time rendering with photorealism. The field of 3D Gaussian-based avatar reconstruction [4], [30], [31] has rapidly developed, becoming a vibrant research area. However, previous works typically rely on clean training data, assuming well-designed human motion in monocular videos without motion blur, which hinders performance in real-world scenarios. This limitation persists regardless of the representation used (see Fig. 4). In contrast, *our approach models the velocity field to localize blur degradation and render dynamic humans from monocular blurry frames using novel rendering losses.*

Deblurring NeRF/3DGS. Various methods [32], [8], [33] have adapted NeRF/3DGS to generate sharp outputs from blurry inputs. Deblur-NeRF [8] was the first to address deblurring within NeRF without requiring clear images during training, using a compact multi-layer perception (MLP) to estimate a per-pixel blur kernel. Deblur-GS [33] applies camera motion deblurring to 3D Gaussian scene representations. However, these methods primarily target static scenes and address blur from camera shifts or defocus. Additionally, recent works [34], [35], [36], [10], [37], [38] have also leveraged event cameras, which are more robust to motion blur, to optimize NeRF/3DGS. *Differently, we are the first to explore event cameras as guidance for rendering dynamic humans in diverse blurry scenes.* Unlike prior methods that roughly integrate RGB and event data, our approach uses both modalities in a complementary manner, tailoring them to human modeling.

Video Deblurring. Deblurring remains a challenging problem in image restoration due to its ill-posed nature. Recent approaches [39], [40], [41] leverage large datasets of blurry and sharp image pairs for supervised learning. Event-based methods [12], [13], [42] have also been explored for motion deblurring, capitalizing on the motion information in events. For human reconstruction from blurry video, these methods provide a *two-stage baseline*, deblurring first and then reconstructing the 3D human. However, most methods struggle with limited generalization across scenes and neglect human-specific priors, often failing with complex motion. Thus, applying them directly to our task leads to errors in appearance and shape in rendered humans (see Fig. 4).

III. Methodology

A. Preliminaries and Background

HumanNeRF [1]. HumanNeRF extends NeRF to render humans from monocular video by modeling them as neural fields. A dynamic human is first represented in a static canonical space, then mapped to the observation space for each pose $\boldsymbol{p}$. The color $\boldsymbol{c}$ and density σ are modeled as:

$$\boldsymbol{c}, \sigma = \mathtt{F}(\hat{\boldsymbol{x}}), \quad \hat{\boldsymbol{x}} = \mathtt{T}(\boldsymbol{x}, \boldsymbol{p}), \tag{1}$$

where $\mathtt{F}$ maps the point $\hat{\boldsymbol{x}}$ in canonical space to color and density, and $\mathtt{T}$ maps the observation space point $\boldsymbol{x}$ to the canonical coordinates $\hat{\boldsymbol{x}}$. Volumetric rendering is applied by aggregating the color and density of the sampled points $\{\boldsymbol{x}_i\}_{i=1}^{P}$ along the ray $\boldsymbol{r}$: $\boldsymbol{C}(\boldsymbol{r}) = \sum_{i=1}^{P} T_i \alpha_i \boldsymbol{c}_i$, where $T_i = \exp(-\sum_{j=1}^{i-1} \alpha_j \delta_j)$ and $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$ represent transmittance and alpha for each sample $\boldsymbol{x}_i$, with δ_i being the interval between samples.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

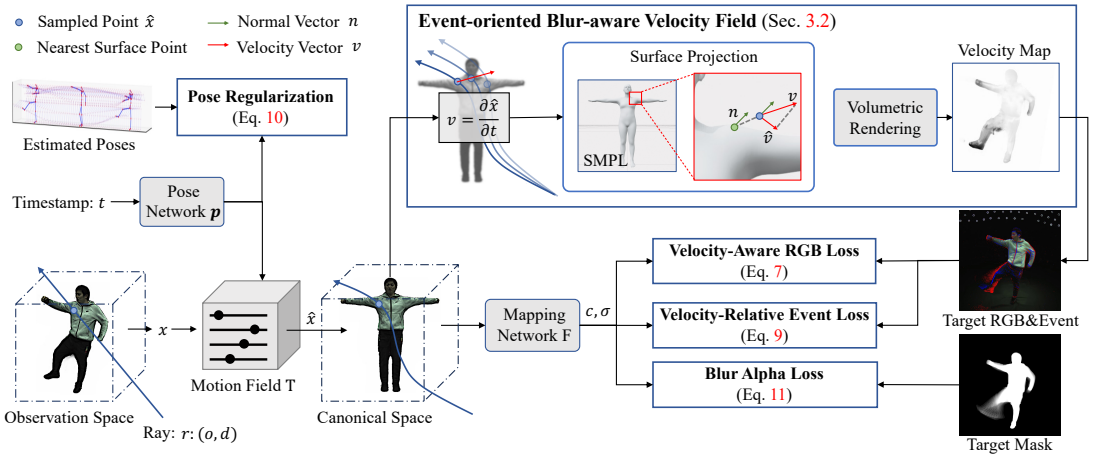

Fig. 2. Overview of our ExFMan: Since our method is based on a per-scene optimization pipeline, the inputs serve as supervision during the optimization phase within the loss function. The RGB loss and Alpha loss incorporate RGB and event data, respectively. The Alpha loss uses the human mask, obtained from the RGB data through an off-the-shelf segmentation model [22], while pose regularization relies on human poses extracted from events using an existing pose estimation model [18].

Event Representation. Events are triggered when intensity changes at a pixel exceed a threshold, offering low latency and high dynamic range. Let $\mathbf{I}_{u,v}(t)$ be the intensity at pixel (u, v) at time t, and $\mathbf{L}_{u,v}(t)$ its logarithm. An event with polarity $p = \pm 1$ is triggered when $\Delta\mathbf{L}_{u,v}(t)$ surpasses a threshold Θ, where polarity indicates the direction of the change. Following prior work [35], [10], we define the integral of events at pixel (u, v) over time interval $(t_i, t_j]$ as: $\hat{E}_{u,v}(t_i, t_j) = \sum_{(t,p):t_i<t\leq t_j} p$, which accounts for the effective number of events in this interval. Θ is fixed and symmetric with respect to polarity.

B. Event-oriented Blur-aware Velocity Field

Consistency across frames is vital for accurate human reconstruction, but motion blur often causes the mapping network F to produce indistinct color and density, leading to inconsistencies. To mitigate this, this module is proposed to reduce the influence of blurry regions and refine them with event data to recover details. Previous methods [43], [44] deploy uncertainty estimation to detect transient regions causing view inconsistency and ambiguity. They represent the radiance values of a scene with a Gaussian distribution, treating the predicted uncertainty as the distribution's variance. Minimizing the Gaussian distribution's negative log-likelihood leads to a scenario where a pixel with high uncertainty is assigned with low influence in the reconstruction. Ideally, pixels with motion blur should be allocated a high variance to lessen their impact on the reconstruction. However, empirical study reveals that the implicit uncertainty fails to precisely characterize the regions with motion blur (Fig. 3.b). We contend that the human representation with complex deformation *overfits* the input blurry frames, yielding a lower uncertainty.

To address this issue, we explicitly formulate the uncertainty related to human motion across the multi-frame context to identify the motion blur. Intuitively, motion blur occurs when the pixels observe multiple regions of humans during the exposure interval. Theoretically, based on the human representation in HumanNeRF [1], motion blur occurs when the canonical point $\hat{\boldsymbol{x}}$ corresponding to the observation point $\boldsymbol{x}$ moves in the exposure interval. Thus, it is reliable to represent the motion blur by employing the velocity of the canonical point $\hat{\boldsymbol{x}}$ *w.r.t* timestamp t. The large velocity means that the observation point moves fast in the canonical space, which naturally yields motion blur. To achieve this, we formulate a 3D velocity field for the human body in canonical space. As is shown in Fig. 2, by modeling the pose as a function of timestamp t, the velocity field $\boldsymbol{v}(\boldsymbol{x}; t)$ can be directly calculated as the derivative of the canonical point $\hat{\boldsymbol{x}}$ *w.r.t* timestamp t,

$$\boldsymbol{v}(\boldsymbol{x}; t) = \frac{\partial \hat{\boldsymbol{x}}}{\partial t} = \frac{\partial \mathtt{T}(\boldsymbol{x}, \boldsymbol{p}(t))}{\partial t}, \tag{2}$$

where $\boldsymbol{p}(t)$ is realized using MLP to optimize continuous human poses. In practice, for convenience and fast optimization, we employ finite differences to estimate derivatives:

$$\boldsymbol{v}(\boldsymbol{x}; t) \approx (\mathtt{T}(\boldsymbol{x}, \boldsymbol{p}(t + \Delta t)) - \mathtt{T}(\boldsymbol{x}, \boldsymbol{p}(t)))/\Delta t, \tag{3}$$

where Δt is a constant time interval.

As shown in Fig. 3 (Right), for the sampled observation point $\boldsymbol{x}$, deformation mapping T deforms it to canonical points of $\hat{\boldsymbol{x}}_t$ and $\hat{\boldsymbol{x}}_{t+\Delta t}$ corresponding to time t and $t + \Delta t$ respectively. Additionally, we observe that although some regions capture similar motion strength, they exhibit varying degrees of blur. The direction of motion plays an important role in this difference. Specifically, for the same motion strength, motion perpendicular to the camera's optical axis results in greater blur than motion that is horizontal to the optical axis. When projected onto the human surface to formulate the velocity field, motion blur is mostly caused by motion with velocity parallel to the human surface. To model this, we decompose the initial velocity into normal and tangential components relative to the surface and use the tangential component to precisely represent the motion blur. Formally, we project the velocity based on the unitary normal vector $\boldsymbol{n}$ of the nearest vertice on the SMPL model [45] in the canonical space, as is shown in Fig. 2 (*red rectangle*),

$$\hat{\boldsymbol{v}}(\boldsymbol{x}; t) = proj_{\boldsymbol{n}}(\boldsymbol{v}(\boldsymbol{x}; t)), \tag{4}$$

where $proj_{\boldsymbol{n}}(\boldsymbol{v}) = \boldsymbol{v} - (\boldsymbol{v} \cdot \boldsymbol{n})\boldsymbol{n}$ projects the vector $\boldsymbol{v}$ onto the plane perpendicular to the normal vector $\boldsymbol{n}$, where $\cdot$ indicates the dot product. Finally, velocity $V(\boldsymbol{r}; t)$ in the image space

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

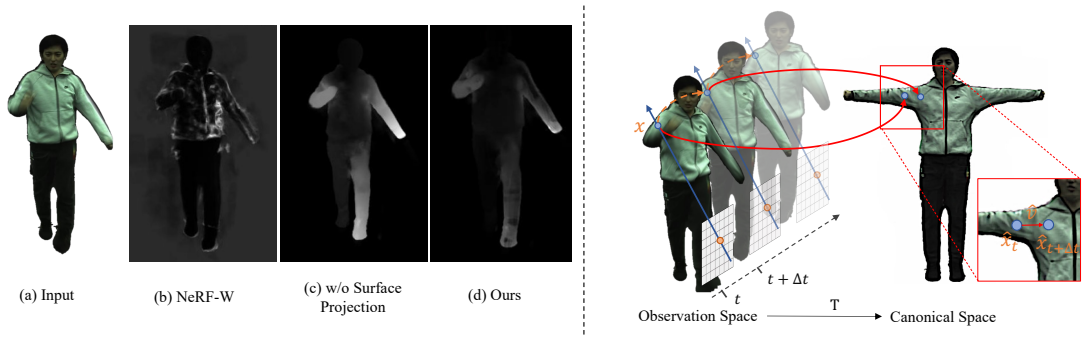

Fig. 3. **Left**: Comparison of velocity field implementations. (b) NeRF-W fails to capture meaningful velocity with implicit uncertainty. (c) Velocity without surface projection highlights incorrect regions. **Right**: Illustration of velocity computation for human model in rapid motion.

is estimated via volumetric rendering,

$$V(\boldsymbol{r};t) = \sum_{i=1}^{P} T_i \alpha_i \overline{v}(\boldsymbol{x}_i;t), \\ \overline{v}(\boldsymbol{x}_i;t) = v_0 + \log(1+\exp(||\hat{\boldsymbol{v}}(\boldsymbol{x}_i;t)||)), \tag{5}$$

where v_0 ensures a minimum variance for all the field; $||*||$ obtains the norm of the velocity vector. The effectiveness with surface projection of Eq. 4 is illustrated in Fig. 3(d), where the velocity $V(\boldsymbol{r};t)$ effectively localizes the region with motion blur. Without surface projection (Fig. 3(c)), the velocity highlights the left arm, but it is less blurry than the right hand. This reflects that the velocity on the left hand without surface projection is mostly vertical to the surface.

C. Velocity-Based Rendering Losses

Velocity-Aware Photometric Loss. Drawing inspiration from NeRF-W [43], we also model the color rendering of a ray using a Gaussian distribution $\overline{\boldsymbol{C}}(\boldsymbol{r}) \sim (\boldsymbol{C}(\boldsymbol{r}), \beta^2(\boldsymbol{r}))$, where $\boldsymbol{C}(\boldsymbol{r})$ represents the mean and $\beta^2(\boldsymbol{r})$ denotes the variance. $\boldsymbol{C}(\boldsymbol{r})$ is determined with rendering result of HumanNeRF. The variance $\beta^2(\boldsymbol{r})$ is explicitly represented with the estimated velocity map $V(\boldsymbol{r})$ in Eq. 5. Following NeRF-W and ActiveNeRF [44], we enhance the photometric loss by minimizing the negative log-likelihood of the distribution of rays $\boldsymbol{r}$ from a batch $\mathbb{R}$,

$$\boldsymbol{L}_{RGB} = \sum_{\boldsymbol{r}\in\mathbb{R}} \frac{||\boldsymbol{C}(\boldsymbol{r}) - \hat{\boldsymbol{C}}(\boldsymbol{r})||}{2V(\boldsymbol{r})} + \frac{\log V(\boldsymbol{r})}{2}, \tag{6}$$

where t is omitted for simplicity, and $\hat{\boldsymbol{C}}(\boldsymbol{r})$ denotes the ground truth color of camera ray $\boldsymbol{r}$. Given that $V(\boldsymbol{r})$ is determined explicitly through Eq. 5 rather than being learned implicitly, it remains constant since we assume the estimated human pose is precise and unchanged. Thus the $V(\boldsymbol{r})$ can be excluded from the second term of the objective function. Additionally, following prior arts [1], [7] for human reconstruction, we also employ a combination of MSE loss and LPIPS [46] loss. Consequently, we define our velocity-aware photometric loss as follows:

$$\boldsymbol{L}_{RGB} = \sum_{\boldsymbol{r}\in\mathbb{R}} \frac{||\boldsymbol{C}(\boldsymbol{r}) - \hat{\boldsymbol{C}}(\boldsymbol{r})||}{2V(\boldsymbol{r})} \\ + \lambda \sum_{\boldsymbol{p}\in\mathbb{P}} \frac{||\mathtt{M}(\boldsymbol{C}(\boldsymbol{p})) - \mathtt{M}(\hat{\boldsymbol{C}}(\boldsymbol{p}))||}{2V(\boldsymbol{p})}, \tag{7}$$

where $\boldsymbol{p}$ is the sampled patch for the perception model $\mathtt{M}$ of LPIPS, λ is the weight balance coefficient.

Velocity-Relative Event Loss. While the velocity-aware photometric loss mitigates inconsistencies in regions affected by motion blur, the regions (*e.g.* , hands) frequently subject to such blur are always assigned high velocity and remain under-constrained. Following previous works [35], [10], [47], the event loss is based on the event integral to supervise the two predicted frames corresponding to the sampled timestamps. However, directly applying the event loss can lead to conflicts in areas well-reconstructed with RGB frames, as noise is often present in events. To address this, we introduce a novel velocity-relative event loss to provide adaptive constraints for these regions. The key idea is to apply a relative weight to the event loss based on the velocity value, allowing for targeted supervision of under-constrained regions using the information from well-reconstructed regions.

Given a pixel (u, v), two moments $\{t_i, t_j\}$ are randomly sampled from in the exposure interval of a frame. We take the difference of the predict colors $\{\boldsymbol{C}_{t_i}, \boldsymbol{C}_{t_j}\}$ in the log domain and divide it by the threshold Θ. Then, an estimate of the number of events between two frames for the pixel is obtained:

$$E_{u,v}(\boldsymbol{C}_{t_i}, \boldsymbol{C}_{t_j}) = \frac{log\boldsymbol{C}_{t_i} - log\boldsymbol{C}_{t_j}}{\Theta}, \tag{8}$$

where $\boldsymbol{C}_{t_*} = \boldsymbol{C}(\boldsymbol{r}_{t_*})$. For simplicity, the ray $\boldsymbol{r}$ corresponding to pixel (u, v) is omitted. To effectively constrain regions affected by motion blur, we introduce a velocity-relative event loss by *anchoring high-velocity pixels to those with low velocity*, as follows:

$$\boldsymbol{L}_{Event} = \sum_{\boldsymbol{r}\in\mathbb{R}} \beta \mathbb{I}_{\beta>1} ||E_{u,v}(\boldsymbol{C}_{t_i}, st(\boldsymbol{C}_{t_j})) - \hat{E}_{u,v}(t_i, t_j)||^2 \\ + \frac{1}{\beta}(1 - \mathbb{I}_{\beta>1}) ||E_{u,v}(st(\boldsymbol{C}_{t_i}), \boldsymbol{C}_{t_j}) - \hat{E}_{u,v}(t_i, t_j)||^2, \tag{9}$$

where $\beta = V_{t_i}/V_{t_j}$ and $\mathbb{I}_{\beta>1}$ indicates if the β is > 1, and $st(*)$ stop the gradient of the regions with low velocity for optimization. $\hat{E}$ denotes the ground true event integral. While event data in our framework serves as a supplementary modality to RGB, its utilization is reflected in two key aspects of our approach. First, in Eq. 9, the timestamps t_i and t_j are specifically adjusted within a smaller interval compared to RGB frames, considering the high temporal resolution of the event camera. This adjustment enables the optimization of finer details, such as the textures of the human body in motion. Second, event data also influences the optimization of the pose network, assisting in the capture of continuous

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I

QUANTITATIVE COMPARISON BETWEEN OUR EXFMAN AND RELATED METHODS ON THE ZJU-MOCAP [2] DATASET. WE COLOR CELLS THAT HAVE THE BEST METRIC VALUES AND SECOND ONES. NOTE THAT LPIPS*=LPIPS $\times 10^3$.

Method		Subject **377**			Subject **386**			Subject **387**		
		PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
Baselines	HumanNeRF [1]	18.66	0.9463	58.50	22.92	0.9650	42.36	15.45	0.9111	113.87
RGB-based Deblur	MPR [11]-HumanNeRF [1]	19.05	0.9479	54.92	23.08	0.9657	40.68	19.74	0.9414	61.96
	MPR [11]-MonoHuman [7]	19.72	0.9510	54.86	23.16	0.9649	45.78	20.00	0.9428	59.78
	MPR [11]-GauHuman [4]	20.42	0.9487	46.70	23.06	0.9585	41.01	20.30	0.9349	58.84
RGB+Event Deblur	D2Net [13]-HumanNeRF [1]	18.47	0.9452	59.51	22.31	0.9623	45.70	17.32	0.9279	78.73
	EFNet [42]-HumanNeRF [1]	18.41	0.9445	55.88	20.66	0.9545	63.80	15.65	0.9088	112.61
	Ours (ExFMan)	23.80	0.9676	38.61	24.72	0.9684	50.46	22.59	0.9493	65.08

Method		Subject **392**			Subject **393**			Subject **394**		
		PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓	PSNR↑	SSIM↑	LPIPS*↓
Baselines	HumanNeRF [1]	16.51	0.9297	81.06	16.90	0.9250	82.41	17.36	0.9326	72.54
RGB-based Deblur	MPR [11]-HumanNeRF [1]	16.85	0.9318	78.07	18.11	0.9329	72.97	18.48	0.9387	65.92
	MPR [11]-MonoHuman [7]	18.44	0.9398	73.21	19.04	0.9379	68.81	19.43	0.9430	63.79
	MPR [11]-GauHuman [4]	18.12	0.9327	68.50	18.91	0.9285	68.03	19.74	0.9367	57.71
RGB+Event Deblur	D2Net [13]-HumanNeRF [1]	16.16	0.9263	85.79	16.76	0.9244	84.33	17.30	0.9308	75.55
	EFNet [42]-HumanNeRF [1]	16.28	0.9267	80.32	16.99	0.9263	76.40	17.62	0.9337	67.37
	Ours (ExFMan)	22.98	0.9600	59.72	22.20	0.9498	66.98	22.62	0.9550	52.73

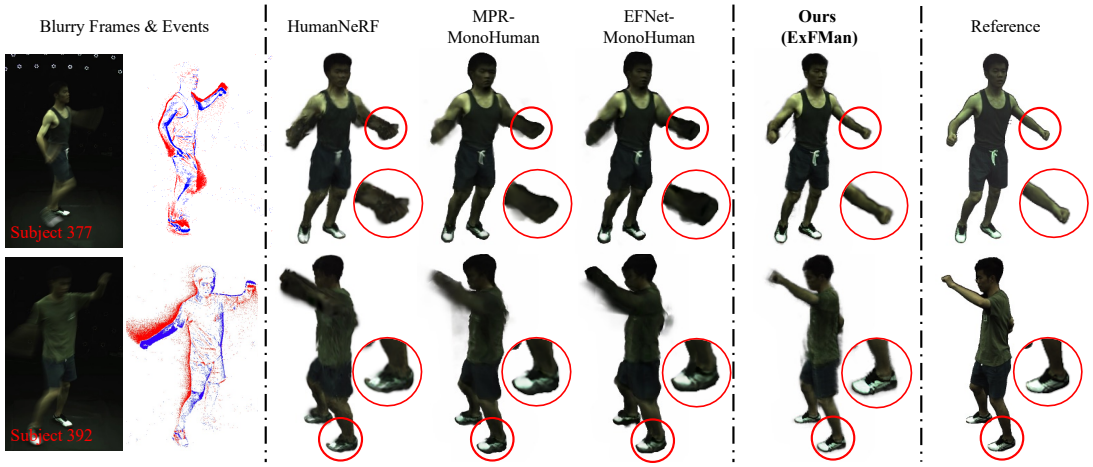

Fig. 4. Qualitative results of novel view synthesis in the ZJU-MoCap dataset [2].

poses through gradient back-propagation. This results in a more precise estimation of the velocity field and contributes to the effectiveness of our uncertainty-aware framework.

D. Optimization

Pose Regularization. To formulate the velocity as a function of timestamp, the time-to-pose network is jointly trained with the human. To incorporate the human prior to constrain the pose network, we add the pose regularization,

$$\boldsymbol{L}_{Pose} = ||\boldsymbol{p}(t_i) - \hat{\boldsymbol{p}}(t_i)|| \tag{10}$$

where $\hat{\boldsymbol{p}}(t)$ denotes the ground true pose from the frame of the exposure timestamp t, which is obtained by manual annotations or a pre-trained estimator (see Sec. IV).

Velocity-based Alpha Loss. Following NeuMan [26], we apply regularization to ensure the accumulated alpha map from the human model aligns with the detected human mask. However, since the masks of blurry frames provided by the matting model [22] are also blurred, directly optimizing with them may introduce boundary ambiguity. To address this, we extend the alpha loss by incorporating the velocity of the human body, allowing for adaptive optimization of a clearer human boundary:

$$\boldsymbol{L}_{Alpha} = \sum_{\boldsymbol{r}\in\mathbb{R}} \frac{||A(\boldsymbol{r}) - \hat{A}(\boldsymbol{r})||}{2V(\boldsymbol{r})}, \tag{11}$$

where A represents the predicted mask obtained through volumetric rendering of α_i in the neural field, indicating opacity with a value of 1 and transparency with a value of 0. $\hat{A}$ denotes the estimated alpha mask from pre-trained models. Based on the velocity-based alpha loss, the human mask is optimized with a focus on regions with clear alpha edges, effectively filtering out the blurry alpha values.

Overall Objective. Incorporating the photometric loss of Eq. 7 and event loss of Eq. 9, the overall objective is formulated to optimize the human model,

$$\boldsymbol{L} = \boldsymbol{L}_{RGB} + \alpha_e \boldsymbol{L}_{Event} + \alpha_p \boldsymbol{L}_{Pose} + \alpha_a \boldsymbol{L}_{Alpha}, \tag{12}$$

where α_e, α_p and α_a are the are balancing weights.

IV. EXPERIMENTS

A. Implementation Details

Datasets: We conduct on two datasets: **1)** Synthetic Data: We extend the ZJU-MoCap dataset [2] with simulated motion blur and event data, focusing on six subjects. The training set uses camera 1, and the remaining 22 cameras serve for

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

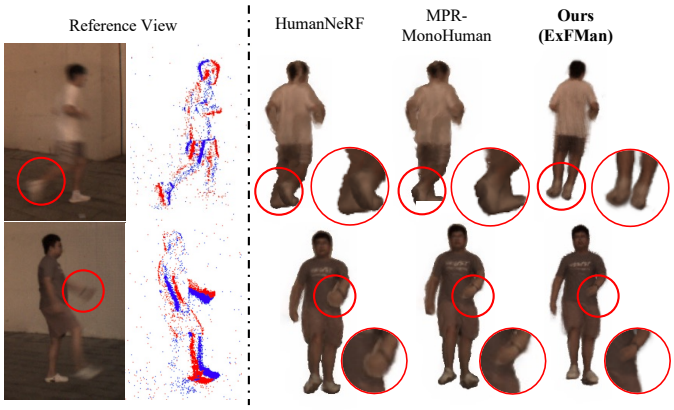

Fig. 5. Qualitative results of novel view synthesis in our real-world dataset.

TABLE II
QUANTITATIVE COMPARISON ON THE REAL-WORLD DATASET. BRISQUE [53] IS UTILIZED FOR EVALUATION (*less is better*).

	Subject **1**	Subject **2**	Subject **3**	Subject **4**
HumanNeRF [1]	75.00	77.25	82.55	80.23
MPR [11]-MonoHuman [7]	74.58	76.47	82.74	79.89
Ours (ExFMan)	59.77	63.70	68.66	65.45

TABLE III
ABLATION STUDY ON THE VELOCITY FIELD, WHERE "VA-PL" INDICATES VELOCITY-AWARE PHOTOMETRIC LOSS AND "VR-EL" INDICATES VELOCITY-RELATIVE EVENT LOSS. THE RESULTS ARE CONDUCTED ON SUBJECT 377 OF ZJU-MOCAP [2].

Method	VA-PL	VR-EL	PSNR↑	SSIM↑	LPIPS*↓
HumanNeRF [1]	-	-	18.66	0.9463	58.50
Baseline	✗	✗	20.68	0.9568	43.55
Baseline w/ VA-PL	✓	✗	22.94	0.9665	41.42
Baseline w/ VR-EL	✗	✓	21.52	0.9615	43.33
ExFMan (full)	✓	✓	**23.80**	**0.9676**	**38.61**

validation. To simulate motion blur, we interpolate seven additional frames between consecutive ones using the RIFE [48], producing blurry frames. Event data is generated with the V2E [49]. Poses and segmentation masks are averaged from sharp images. **2)** Real-World Data: We use the DAVIS346 color event camera [50], capturing spatial-temporally aligned events and RGB frames at a resolution of 346×260 and an RGB exposure time of 100 ms. The camera is stationary while the subject performs rapid motions (*e.g.*, running, jumping), causing motion blur in the RGB frames. We estimate human poses using EventHPE [18], refining them with SMPLify [51] and OpenPose [52], and obtain segmentation masks with RVM [22].

Implementation Details. We train each scene for 200k iterations on an NVIDIA A800 GPU. We set $\alpha_e = 0.2$, $\alpha_p = 0.01$, and $\alpha_a = 0.01$ with a batch size of 12 patches (20×20 each). For event optimization, we set the threshold Θ to 0.2, following E2NeRF [10]. Our framework consists of three primary trainable networks: the mapping network, the motion field network, and the pose network. The mapping network utilizes an 8-layer MLP with a width of 256, which takes positional encodings as input and outputs color and density. The motion field network, inspired by the approach in HumanNeRF [1], starts with a fully connected layer followed by five convolutional layers. The pose network uses a 4-layer MLP with a width of 256. In inference, the framework requires 8 GB of memory with a batch size of 1 and processes 4 seconds per frame, which is comparable to HumanNeRF.

Compared Methods. We compare ExFMan with HumanNeRF [1], MonoHuman [7] and GauHuman [4]. As no methods directly utilize events for this task, we use a two-stage pipeline: deblurring videos, then reconstructing the human. For deblurring, we use RGB-based methods (MPR [11]) and RGB+Event methods (D2Net [13], EFNet [42]).

B. Results on Synthesis Dataset

We evaluate the models using PSNR, SSIM, and LPIPS [46]. As shown in Tab. I, ExFMan outperforms the baselines, especially in PSNR and SSIM. While deblurring methods like MPR-MonoHuman show improvements in certain metrics, their overall performance is limited due to inconsistent human modeling during deblurring, resulting in poor generalization. For subject 386, ExFMan achieves the highest PSNR and SSIM, while the comparison methods slightly outperform it in LPIPS due to slower motion and less blur. Among the two-stage deblurring methods, MPR-MonoHuman excels in PSNR and SSIM, while MPR-GauHuman performs better in LPIPS. However, both NeRF- and 3DGS-based methods suffer from similar degradation, highlighting the challenges of the motion blur in the dataset. Qualitatively, Fig. 4 shows that ExFMan consistently recovers details, especially in motion-heavy regions like hands (see Subject 392). In contrast, the HumanNeRF baseline loses detail and produces vague boundaries, while two-stage methods MPR-MonoHuman and EFNet-MonoHuman struggle with fine details due to a lack of temporal consistency. ExFMan optimizes the human model end-to-end, ensuring global consistency and improved detail recovery in motion blur-affected regions.

C. Results on Real-World Dataset

Due to the lack of multi-view ground truth sharp images in real-world data, we use the no-reference image quality metric BRISQUE [53] for quantitative analysis. We synthesize 360-degree renderings from 5 frames with 100 views, and ExFMan achieves the best results, outperforming the MPR deblurring method, as shown in Tab. II. Qualitative results (Fig. 5) show that the baseline method (*i.e.*, HumanNeRF) produces hazy boundaries, while the two-stage MPR-MonoHuman shows only slight improvements but still fails to recover clear boundaries. In contrast, ExFMan preserves color information and sharp human boundaries across all subjects. For example, in the 1^{st} row, it clearly separates the legs, while other methods merge them.

D. Ablation Study and Analysis

Impact of the velocity field. To reveal the effectiveness of the velocity field, we conduct an ablation study in Tab. III, where "VA-PL" indicates velocity-aware photometric loss and "VR-EL" indicates velocity-relative event loss. For the *baseline* method, without a velocity field, we optimize the vanilla photometric loss and event loss, alongside other regularization loss (*i.e.*, pose regularization and velocity-based alpha loss). It is shown that velocity-aware photometric loss

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

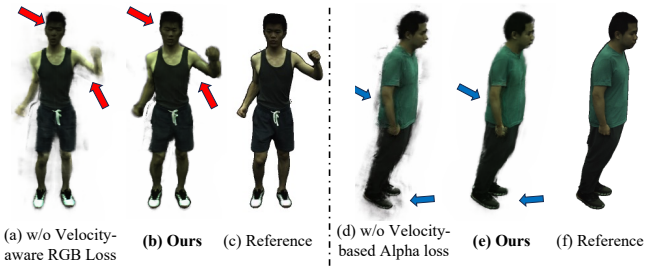

Fig. 6. **Left:** Velocity-aware photometric loss improves rendering quality at the region with motion blur ((a) vs. (b)). **Right:** The velocity-based alpha loss obtain clearer boundary compared with baseline ((d) vs. (e)).

obtained significant improvement compared with the baseline. We also show the rendered result of the comparison on photometric loss in Fig. 6 (left), where blur and haze are exhibited on the regions of arms for the baseline method.

TABLE IV
SENSITIVE STUDY FOR α_e

α_e	PSNR↑	SSIM↑	LPIPS*↓
0	23.06	0.9589	48.70
0.02	23.32	0.9639	47.27
0.1	24.01	0.9670	41.62
0.2	23.80	0.9676	38.61

TABLE V
SENSITIVE STUDY FOR α_a

α_a	PSNR↑	SSIM↑	LPIPS*↓
0	23.34	0.9662	50.43
0.001	24.04	0.9664	49.47
0.005	23.97	0.9667	42.74
0.01	23.80	0.9676	38.61

It is believed that velocity helps to reduce the ambiguity caused by motion blur in the photometric loss and improves the performance. Moreover, velocity-relative event loss also contributes to the performance with supplemental data constrain for the blurry regions, while the full model achieves the best performance.

Impact of the velocity-based alpha loss. The velocity-based alpha loss is designed to handle motion blur by encouraging the rendered alpha to focus on the clear human mask, taking into account human motion. As shown in Fig. 6 (right), our method achieves clearer boundaries compared to the vanilla alpha loss, while the baseline method tends to produce artifacts around the human subject. These artifacts are a result of optimizing for the blurry human masks.

Impact of event data. We conducted experiments to evaluate the role of event data by comparing it with a baseline that excludes it. This variant optimizes only the velocity-aware photometric and regularization terms, yielding a PSNR of 23.06 dB and LPIPS of 48.70 on subject 377. Incorporating event data improved performance by **0.74 dB** in PSNR and **10.09** in LPIPS, highlighting its importance in enhancing human reconstruction amidst motion blur.

Impact of motion blur intensity. We recorded human motion during running with varying exposure times, resulting in different levels of motion blur. In Fig. 7, we analyze how the performance of the proposed approach changes with varying motion blur intensity. By increasing running speed and exposure time, we can increase the blur intensity. Notably, even as the blur intensity increases, the details of appearance and texture remain clear. However, when the motion speed becomes excessively fast, it exceeds the limits of both the human body and the capture system.

Sensitivity study for balancing weights. In the overall loss, the balancing weights are set to adjust the loss terms to a similar magnitude. We conducted further experiments on

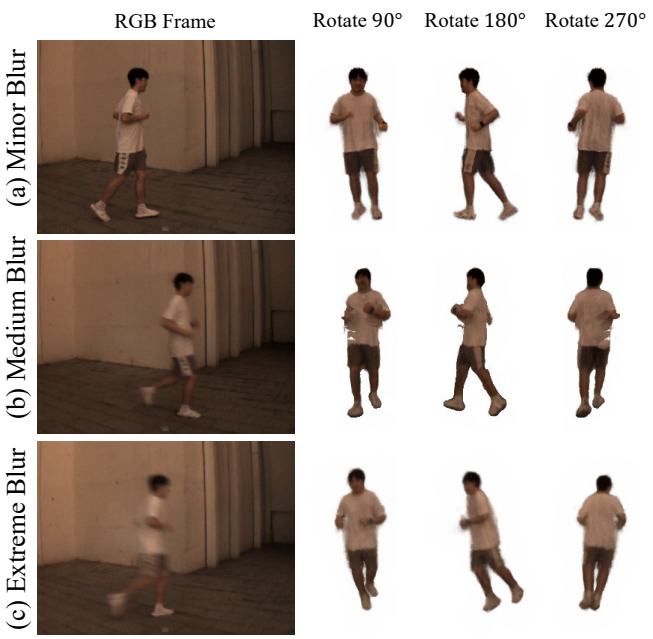

Fig. 7. Impact of Motion Blur Intensity. We set exposure times to 20 ms, 40 ms, and 60 ms, corresponding to minor, medium, and extreme levels of motion blur. Our method demonstrates robustness across varying degrees of motion blur intensity.

subject 377 with different balancing weights to evaluate their effects. For α_e in Tab. IV, we observed that while the PSNR fluctuates within a small range, the best SSIM and LPIPS values are achieved when $\alpha_e = 0.2$. For α_a in Tab. V, we achieve the best LPIPS when $\alpha_a = 0.01$. Intuitively, the event data and alpha mask are useful in retaining structural information, such as body boundaries, which benefits LPIPS. For α_p, the values fluctuate within a reasonable range. Since the pose regularization facilitates the optimization of the small pose network (4-layer MLP), it converges easily.

V. DISCUSSION AND CONCLUSION

Discussion. Although our method shows promising results in human reconstruction under motion blur, several limitations remain. First, our framework is currently applied to static scenes with a static monocular camera. Handling dynamic scenes with dynamic cameras introduces additional challenges, which remains an open area for future research. Additionally, estimating the velocity field increases training computational costs. This can be addressed with advanced human representations (e.g., hash grids [23], 3D Gaussian Splatting [6]). Moreover, ExFMan may produce subtle artifacts at the boundaries. This is due to our reliance on pose and mask estimates from pre-trained models, which can be affected by extreme motion blur.

Conclusion. We propose ExFMan to address the challenge of rendering humans from motion blur, a task where previous methods struggle due to inconsistencies in shape and appearance. We incorporate hybrid data of blurry RGB frames and events, and design a framework based on a novel velocity field to facilitate rendering in motion blur. Our experiments on synthetic and real-world datasets show that ExFMan outperforms baselines and state-of-the-art methods in terms of appearance recovery and shape reconstruction.

REFERENCES

[1] C.-Y. Weng, B. Curless, P. P. Srinivasan, J. T. Barron, and I. Kemelmacher-Shlizerman, "HumanNeRF: Free-viewpoint rendering of moving people from monocular video," in *Proc. of CVPR*, 2022.

[2] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proc. of CVPR*, 2021.

[3] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao, "Animatable neural radiance fields for modeling dynamic human bodies," in *Proc. of ICCV*, 2021.

[4] S. Hu, T. Hu, and Z. Liu, "GauHuman: Articulated gaussian splatting from monocular human videos," in *Proc. of CVPR*, 2024.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, 2021.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Transactions on Graphics*, 2023.

[7] Z. Yu, W. Cheng, X. Liu, W. Wu, and K.-Y. Lin, "MonoHuman: Animatable human neural field from monocular video," in *Proc. of CVPR*, 2023.

[8] L. Ma, X. Li, J. Liao, Q. Zhang, X. Wang, J. Wang, and P. V. Sander, "Deblur-NeRF: Neural radiance fields from blurry images," in *Proc. of CVPR*, 2022.

[9] D. Lee, J. Oh, J. Rim, S. Cho, and K. M. Lee, "ExBluRF: Efficient radiance fields for extreme motion blurred images," in *Proc. of ICCV*, 2023.

[10] Y. Qi, L. Zhu, Y. Zhang, and J. Li, "E2NeRF: Event enhanced neural radiance fields from blurry images," in *Proc. of ICCV*, 2023.

[11] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. of CVPR*, 2021.

[12] L. Pan, C. Scheerlinck, X. Yu, R. Hartley, M. Liu, and Y. Dai, "Bringing a blurry frame alive at high frame-rate with an event camera," in *Proc. of CVPR*, 2019.

[13] W. Shang, D. Ren, D. Zou, J. S. Ren, P. Luo, and W. Zuo, "Bringing events into video deblurring with non-consecutively blurry frames," in *Proc. of ICCV*, 2021.

[14] F. Xu, L. Yu, B. Wang, W. Yang, G.-S. Xia, X. Jia, Z. Qiao, and J. Liu, "Motion deblurring with real events," in *Proc. of ICCV*, 2021.

[15] J. Liang, Y. Yang, B. Li, P. Duan, Y. Xu, and B. Shi, "Coherent event guided low-light video enhancement," in *Proc. of ICCV*, 2023, pp. 10 615–10 625.

[16] K. Chen, G. Liang, H. Li, Y. Lu, and L. Wang, "Evlight++: Low-light video enhancement with an event camera: A large-scale real-world dataset, novel method, and more," *arXiv preprint arXiv:2408.16254*, 2024.

[17] G. Liang, K. Chen, H. Li, Y. Lu, and L. Wang, "Towards robust event-guided low-light image enhancement: a large-scale real-world event-image dataset and novel approach," in *Proc. of CVPR*, 2024, pp. 23–33.

[18] S. Zou, C. Guo, X. Zuo, S. Wang, P. Wang, X. Hu, S. Chen, M. Gong, and L. Cheng, "EventHPE: Event-based 3D human pose and shape estimation," in *Proc. of ICCV*, 2021.

[19] L. Xu, W. Xu, V. Golyanik, M. Habermann, L. Fang, and C. Theobalt, "Eventcap: Monocular 3D capture of high-speed human motions using an event camera," in *Proc. of CVPR*, 2020.

[20] J. Jiang, J. Li, B. Zhang, X. Deng, and B. Shi, "EvHandPose: Event-based 3D hand pose estimation with sparse supervision," *arXiv preprint arXiv:2303.02862*, 2023.

[21] J. Nehvi, V. Golyanik, F. Mueller, H.-P. Seidel, M. Elgharib, and C. Theobalt, "Differentiable event stream simulator for non-rigid 3D tracking," in *Proc. of CVPR*, 2021.

[22] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proc. of WACV*, 2022.

[23] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics*, 2022.

[24] C. Gao, A. Saraf, J. Kopf, and J.-B. Huang, "Dynamic view synthesis from dynamic monocular video," in *Proc. of ICCV*, 2021.

[25] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proc. of CVPR*, 2021.

[26] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan, "NeuMan: Neural human radiance field from a single video," in *Proc. of ECCV*, 2022.

[27] H. Lin, S. Peng, Z. Xu, Y. Yan, Q. Shuai, H. Bao, and X. Zhou, "Efficient neural radiance fields for interactive free-viewpoint video," in *SIGGRAPH Asia Conference Proceedings*, 2022.

[28] T. Jiang, X. Chen, J. Song, and O. Hilliges, "InstantAvatar: Learning avatars from monocular video in 60 seconds," in *Proc. of CVPR*, 2023.

[29] C.-Y. Weng, B. Curless, and I. Kemelmacher-Shlizerman, "Vid2Actor: Free-viewpoint animatable person synthesis from video in the wild," *arXiv preprint arXiv:2012.12884*, 2020.

[30] H. Jung, N. Brasch, J. Song, E. Perez-Pellitero, Y. Zhou, Z. Li, N. Navab, and B. Busam, "Deformable 3d gaussian splatting for animatable human avatars," *arXiv preprint arXiv:2312.15059*, 2023.

[31] M. Li, J. Tao, Z. Yang, and Y. Yang, "Human101: Training 100+ fps human gaussians in 100s from 1 view," *arXiv preprint arXiv:2312.15258*, 2023.

[32] Z. Wu, X. Li, J. Peng, H. Lu, Z. Cao, and W. Zhong, "DoF-NeRF: Depth-of-field meets neural radiance fields," in *Proceedings of the ACM International Conference on Multimedia*, 2022.

[33] W. Chen and L. Liu, "Deblur-GS: 3D Gaussian Splatting from Camera Motion Blurred Images," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2024.

[34] I. Hwang, J. Kim, and Y. M. Kim, "Ev-NeRF: Event based neural radiance field," in *Proc. of WACV*, 2023.

[35] V. Rudnev, M. Elgharib, C. Theobalt, and V. Golyanik, "EventNeRF: Neural radiance fields from a single colour event camera," in *Proc. of CVPR*, 2023.

[36] T. Xiong, J. Wu, B. He, C. Fermuller, Y. Aloimonos, H. Huang, and C. A. Metzler, "Event3DGS: Event-based 3D Gaussian Splatting for Fast Egomotion," *arXiv preprint arXiv:2406.02972*, 2024.

[37] M. Cannici and D. Scaramuzza, "Mitigating motion blur in neural radiance fields with events and frames," in *Proc. of CVPR*, 2024.

[38] Z. Zhang, K. Chen, and L. Wang, "Elite-EvGS: Learning Event-based 3D Gaussian Splatting by Distilling Event-to-Video Priors," *arXiv preprint arXiv:2409.13392*, 2024.

[39] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. of CVPR*, 2017.

[40] J. Rim, G. Kim, J. Kim, J. Lee, S. Lee, and S. Cho, "Realistic blur synthesis for learning image deblurring," in *Proc. of ECCV*, 2022.

[41] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *Proc. of ECCV*, 2020.

[42] L. Sun, C. Sakaridis, J. Liang, Q. Jiang, K. Yang, P. Sun, Y. Ye, K. Wang, and L. V. Gool, "Event-based fusion for motion deblurring with cross-modal attention," in *Proc. of ECCV*, 2022.

[43] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the wild: Neural radiance fields for unconstrained photo collections," in *Proc. of CVPR*, 2021.

[44] X. Pan, Z. Lai, S. Song, and G. Huang, "ActiveNeRF: Learning where to see with uncertainty estimation," in *Proc. of ECCV*. Springer, 2022.

[45] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries*, 2023.

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. of CVPR*, 2018.

[47] Q. Ma, D. P. Paudel, A. Chhatkuli, and L. Van Gool, "Deformable neural radiance fields using RGB and event cameras," in *Proc. of ICCV*, 2023.

[48] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *Proc. of ECCV*, 2022.

[49] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic dvs events," in *Proc. of CVPR*, 2021.

[50] G. Taverni, D. P. Moeys, C. Li, C. Cavaco, V. Motsnyi, D. S. S. Bello, and T. Delbruck, "Front and back illuminated dynamic and active pixel vision sensors comparison," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2018.

[51] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. of ECCV*, 2016.

[52] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2019.

[53] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, 2012.