

# 4D E-SloMo: 4D Reconstruction for High Speed Scene using a Hybrid RGB-Event Multi-View System

## Supplementary Material

### 6. Additional Method Details

#### 6.1. Motion-aware Partitioning

In Sec. 6.1, we quantify short-term motion using two complementary measures. First, we define the cumulative brightness change  $a(t)$ , obtained by integrating the denoised event windows from the partition start time  $t_0$  up to time  $t$  at the full sensor resolution ( $1280 \times 720$ ). Second, we compute the cumulative event count  $s(t)$ , defined as the sum of per-frame event counts over the same temporal interval.

To make the two quantities comparable in scale, we normalize  $a(t)$  and  $s(t)$  by factors of  $100K$  and  $5M$ , respectively. A partition is terminated once their sum exceeds unity:

$$\frac{|a(t)|}{200K} + \frac{s(t)}{5M} > 1.$$

The canonical time of each partition is selected by evaluating, for every frame within the partition, its cumulative brightness difference to all other frames' event windows. The frame with the minimum total accumulated difference is chosen as the representative (canonical) event time.

At partition boundaries, we apply *boundary smoothing*. During the late stages of training, Gaussians near the temporal boundary are jointly optimized using supervision from adjacent partitions, together with a temporal total-variation regularization. This encourages a smooth transition of geometry and appearance across partition boundaries.

#### 6.2. Event-to-Flow Pipeline and Flow-based Loss

The event-to-flow pipeline is illustrated in Fig. 5. We first convert raw polarity events into intensity snapshots using an E2VID-style recurrent reconstruction. Optical flow is then estimated between consecutive snapshots. The flow guides the backward warping of the nearest RGB frame to arbitrary event timestamps, enabling a dense temporal supervision signal.

Directly applying the raw occlusion mask obtained from backward warping often results in large invalid regions, which introduces ambiguity in the supervision signal. In extreme cases, moving structures (e.g., arms) may appear at multiple locations due to incorrect warping. To alleviate this issue, we refine the occlusion handling with a conservative background extraction scheme.

We first use per-frame foreground masks to treat pixels that consistently lie outside the foreground across multiple frames as reliable background. Pixels that appear in the

background for more than three occurrences are marked as stable background regions and used to fill the missing areas in the initial occlusion mask. After completing the occlusion map with these background-supported pixels, we take the complement of the refined mask and apply it as a per-pixel weighting term in the photometric loss.

The resulting flow-based photometric loss is expressed as:

$$\mathcal{L}_{\text{flow}} = \left\| (1 - M_{\text{refined}}) \odot (C(t') - \hat{C}_{\text{warp}}(t')) \right\|_1, \quad (10)$$

where  $M_{\text{refined}}$  denotes the completed occlusion mask incorporating background-supported regions.

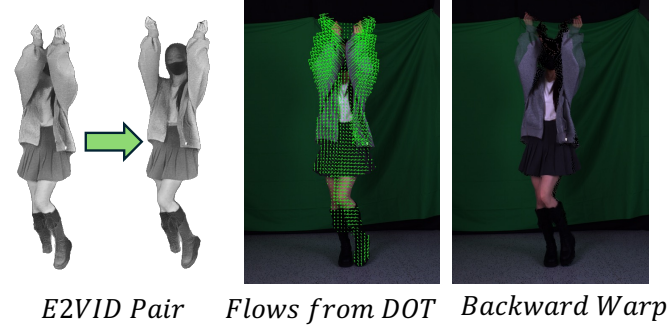


Figure 5. The Event-to-Flow pipeline. We reconstruct an E2VID intensity pair, estimate flows with DOT, and backward-warp the RGB frame. The slight ghosting is only for visualization (a small blend was added highlight the source of the warp); the actual warp does not contain this artifact.

#### 6.3. Event-weighted Sampling

We further design an event-based weighted sampling strategy. In our three-stage optimization pipeline, this module is activated in Stage 2 once a stable geometric initialization has been established. For any given frame, its sampling weight is defined to be proportional to two event-based quantities. Formally, we define

$$w = \lambda_{\text{key}} A_{\text{key}} + \lambda_{\text{nbr}} A_{\text{nbr}}, \quad (11)$$

where  $\lambda_{\text{key}} = 0.1$  and  $\lambda_{\text{nbr}} = 1.0$ ,  $A_{\text{key}}$  denotes the accumulated event window with respect to the keyframe, and  $A_{\text{nbr}}$  denotes the accumulated event window relative to the preceding and succeeding frames.

This weighting scheme assigns higher sampling probabilities to regions with strong motion cues, thereby improving the reconstruction of rapidly deforming structures

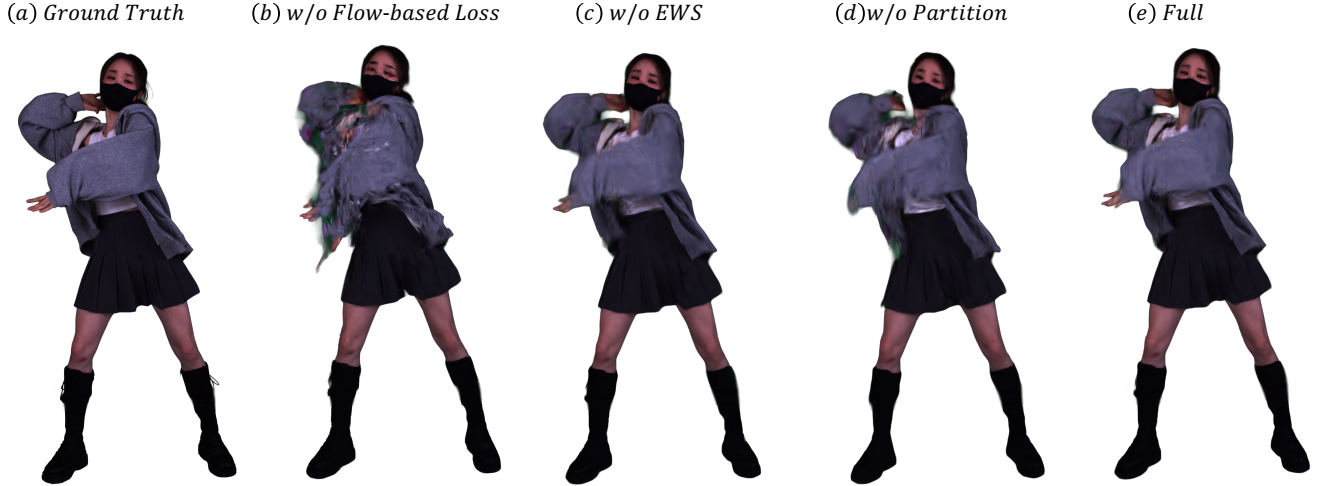


Figure 6. The ablation results on real dataset.

while maintaining the reconstruction quality in most other regions.

#### 6.4. PSNR-ROI Definition

We evaluate quantitative performance using PSNR-ROI (see Sec. 4.5), defined by computing PSNR only within the bounding box of the dynamic foreground region. The foreground mask is produced by applying SAM2[32]-based video segmentation to the ground-truth frames. This ROI formulation focuses the evaluation on motion-dominated regions, where event cues are expected to have the greatest impact.

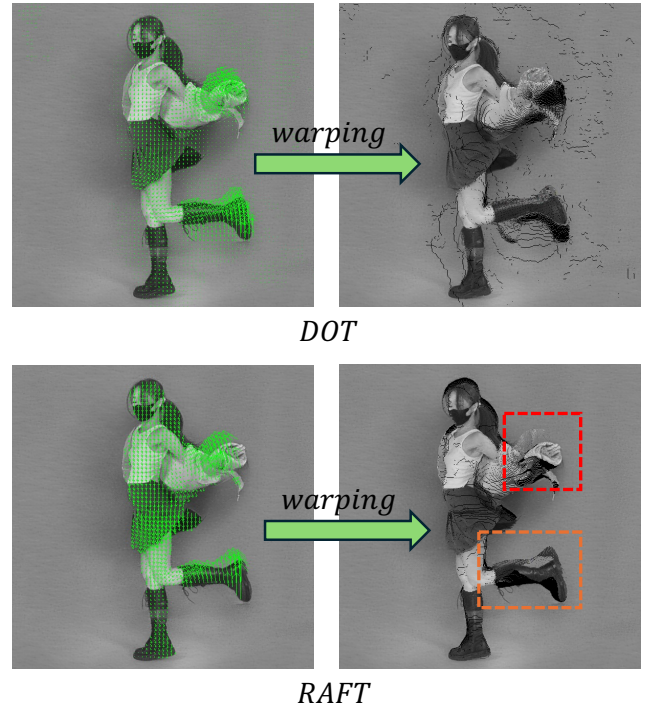
### 7. Additional Experiments and Analysis

#### 7.1. Ablation Visualizations

Fig. 6 presents visual comparisons on the real dataset. Without the flow-based loss (Fig. 6(b)), the model no longer receives motion supervision from events, causing the 4DGS representation to fail at interpolating correct non-rigid motion. Removing the event-weighted sampling module (Fig. 6(c)) leads to degraded reconstruction of fast-moving regions, as the model is no longer guided to focus on frames and areas with strong event activity. When the partition-based training scheme is disabled (Fig. 6(d)), the single-stage optimization struggles with long and complex motions, resulting in smeared geometry and unstable appearance. Our full model (Fig. 6(e)) produces the sharpest structure and most faithful motion reconstruction.

#### 7.2. DOT vs. RAFT Optical Flow

We compare the flow estimation performance of DOT[15] and RAFT[42] under our event-to-flow pipeline (visual results in Fig. 7). The red box and the yellow box respec-

Figure 7. *left*: Flow estimation from DOT[15] and RAFT[42]. *right*: Backward warping results.

tively highlight the regions where warping with RAFT’s optical flow completely failed and partially failed. RAFT produces noticeably less stable flow fields on E2VID reconstructions, particularly under high-speed motion, leading to larger warping errors. DOT, in contrast, better preserves motion discontinuities and handles large displacements more reliably. From algorithmic perspective, DOT

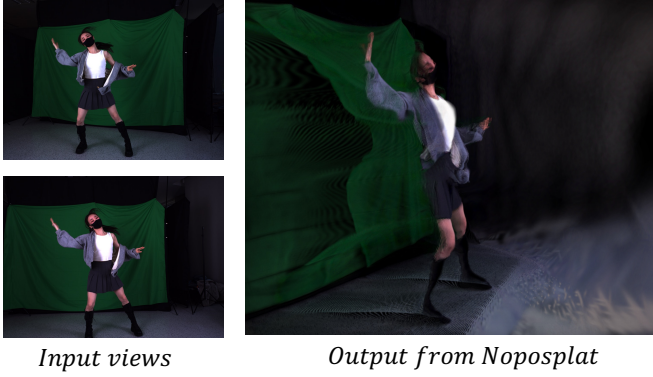


Figure 8. The 3D gaussians decoded from noposplat, used for initialization after uniformly downsampling.

starts with long-range point trajectories that naturally handle large motions and occlusions, providing strong anchor points before the network refines the flow. Because of these trajectory priors, DOT doesn’t have to “blindly search” over a huge correspondence space like RAFT does, making it much more reliable for large displacements.

### 7.3. NoPoSplat Initialization

Due to the sparsity of our camera views, COLMAP often fails to reconstruct a reliable geometry, making it difficult to obtain a usable 3D Gaussian initialization for downstream 4D modeling. To address this, we adopt a feed-forward NoPoSplat model to generate an initial set of 3D Gaussians directly from the available input views. As shown in Fig. 8, even with only 2–3 views, NoPoSplat produces a stable and sufficiently dense point cloud, which serves as an effective starting point for subsequent 4D Gaussian optimization.