1. Method

1.1. Network Architecture

Our method employs a convolutional recurrent neural network designed to reconstruct HDR videos from event streams. As shown in Figure 1, the network processes $T = 2N + 1$ consecutive event voxel grids $\{\mathbf{E}_{t-N}, \ldots, \mathbf{E}_{t+N}\}$ to generate the HDR frame $\mathbf{H}_t$ at timestamp t. The architecture includes several key modules.

Firstly, the shared feature extractor downsamples event frames to a low spatial resolution feature space using strided convolution layers, producing $2N + 1$ output feature maps, $\{\mathbf{F}_{t-N}, \ldots, \mathbf{F}_{t+N}\}$. This shared encoding facilitates subsequent alignment by transforming the input data into a consistent feature space.

We then employ a deformable convolution-based alignment module, which uses pyramidal deformable convolutions to align features of different event frames with the central frame feature $\mathbf{F}_t$. This approach predicts offsets for the convolution kernels through a pyramidal processing structure, allowing the network to handle larger movements and align features accurately, thereby avoiding the pitfalls of inaccurate optical flow estimation.

The aligned features are combined in the attentive fusion and reconstruction module. Here, the features are stacked and processed by attention mechanisms that independently focus on height, width, and temporal/channel correlations. The fused features are passed through a recurrent residual network and a ConvLSTM module, which help maintain temporal continuity by remembering information from successive sequences.

To enhance temporal consistency, we introduce a novel temporal consistency loss based on the integral relationship between consecutive frames and events, modeled using a pre-trained UNet-like network. This loss ensures smooth transitions between frames, mitigating issues related to temporal discontinuity.

1.2. Training Strategy

The training strategy involves a combination of losses to optimize HDR video reconstruction and maintain temporal consistency. Given the reconstructed video sequence $\mathbf{H}_i$ and the corresponding ground truth frames $\hat{\mathbf{H}}_i$, we employ several loss functions.

We use the l_1 loss to measure the pixel-wise difference between the reconstructed and ground truth frames:

$$\mathcal{L}_{l_1} = \sum_{i=1}^{T} \|\mathbf{H}_i - \hat{\mathbf{H}}_i\|. \tag{1}$$

To enhance perceptual quality, we introduce the Learned Perceptual Image Patch Similarity (LPIPS) loss [1], which focuses on high-level and structural similarity. Additionally, the temporal consistency loss $\mathcal{L}_C$, derived from the pre-trained network, is defined as:

$$\mathcal{L}_C = \sum_{i=1}^{T} \|\mathbf{E}_t - \mathcal{C}(\mathbf{H}_{t-1}, \mathbf{H}_t)\|_2^2. \tag{2}$$

This loss ensures that the reconstructed frames maintain smooth transitions.

The overall loss function used to train our model is:

$$\mathcal{L} = \mathcal{L}_{l_1} + \tau_1 \mathcal{L}_{LPIPS} + \tau_2 \mathcal{L}_C, \tag{3}$$

where τ_1 and τ_2 are empirically set to 2 and 0.2, respectively.

The network is initialized using Kaiming initialization, and trained with the Adam optimizer (momentum set to 0.9). The initial learning rate is 10^{-4}, reduced by a factor of 10 every 50 epochs. We set the batch size to 4, and train the model for 100 epochs. The implementation uses the PyTorch framework, and training is performed on NVIDIA TITAN V GPUs.

In summary, our network architecture and training strategy effectively reconstruct high-quality HDR videos from event data, ensuring both spatial and temporal coherence.

References

[1] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 1

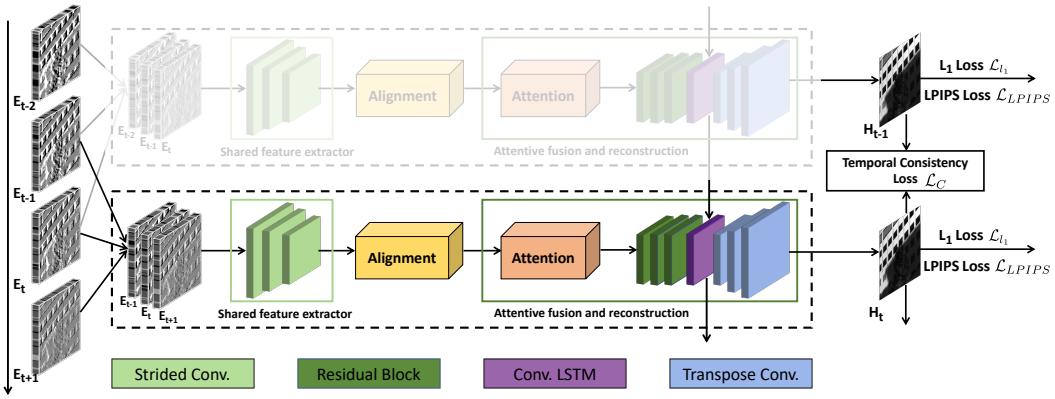

Figure 1. The overview of our recurrent convolutional neural network for HDR video reconstruction from events.