Generating High Dynamic Range Image Sequences with Event Cameras Based on Multi-Head Encoding Networks

1 Introduction

1.1 Introduction to Network

As shown in the figure 1, we introduce three types of inputs at the input end. The first part is data generated from the voxel structure of raw event camera data, with the time step of 8. This part of the data represents the most primitive detail information of the image. The second part is 2D image data generated by e2vid, which serves as the reference frame for LDR 2D images. The third part is HDR image data generated by the decoder, representing the HDR data of historical frames, which can help smooth the entire video and ultimately generate HDR image data for the current frame.

1.2 Introduction to Training Details

According to the aforementioned network architecture, this method is trained using the following approach. Firstly, the training data is converted into voxels based on the original timestamps and stored in the .npy format. For the purpose of facilitating temporal training, each .npy file is stored with a length of T=16. Subsequently, e2vid is employed to generate reference frames corresponding to each .npy file. It is noteworthy that the reference frames generated by e2vid are often affected by the actual data distribution density, leading to occurrences of blank spaces or excessive noise.

Regarding the network itself, data augmentation techniques such as random horizontal flipping and the addition of Gaussian noise with a standard deviation of 0.001 are applied to the input data. Additionally, to better align with the evaluation metrics of this challenge, four types of noise are introduced, including KL divergence noise (to ensure alignment between HDR ground truth images and generated images), L1 noise, L2 noise, and SSIM noise. Due to time constraints, rigorous ablation experiments were not conducted. However, from a holistic analysis of the results, KL divergence noise yielded relatively favorable gains.

Other training parameters include the Adam optimizer with a learning rate of 0.001, cosine annealing for learning rate scheduling, a batch size of 48, and iterative training conducted using four NVIDIA 4090 GPUs. Training is performed for 1000 epochs, with the overall training and testing dataset split in an 8:2 ratio. To address challenging samples, this method manually removes data with poor distributions (some data lack original event information due to bandwidth congestion, rendering them unsuitable for training). Finally, inference can be performed, followed by truncating and normalizing the image with a maximum value of 65535, resulting in the HDR image of the current moment in the corresponding distribution domain.

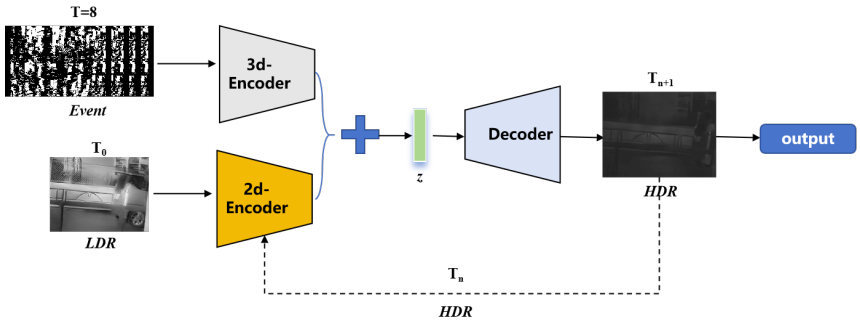

Figure 1: The network architecture.