13th International Conference on Computer and Knowledge Engineering (ICCKE 2023), November 1-2, 2023,
Ferdowsi University of Mashhad, Iran

Intensity-Image Reconstruction Using Event Camera Data by Changing in LSTM Update

Arezoo Rahmati Soltangholi
Dep. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
ar.rahmati@mail.um.ac.ir

Ahad Harati
Dep. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
a.harati @um.ac.ir

Abedin Vahedian
Dep. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
vahedian@um.ac.ir

***Abstract*— Event cameras offer many advantages, but their output is inherently ambiguous and needs to be converted into a more understandable output. One way to use the output of these cameras is to reconstruct the intensity. Various methods have been proposed for image reconstruction using event data, each attempting to improve image quality from specific aspects. In this study, we aim to increase image quality in a challenging condition when the number of events is very low or zero without retraining or changing the network structure during training. Another challenging situation is at the initial start-up moment which requires an initialization time. In this study, we used the potential of the E2VID model and increased the video quality without changing the trained model. Our method performs better than the E2VID method with an 11.9% improvement in the first 10 frames and a 2% improvement in entire videos in SSIM metric**

Keywords*— *event camera, intensity-image reconstruction, deep neural network

I. INTRODUCTION

Event-based cameras are sensors that measure the change in brightness intensity in each pixel asynchronously and operate differently from conventional cameras. They send intensity changes in the form of events. Each event contains information about time, location, and brightness change. The most important advantages of these cameras compared to conventional frame-based ones are high dynamic range, high temporal resolution, and low power consumption. Besides all advantages, using these sensors face challenges. For example, feature extraction in these sensors differs from normal frames because these sensors only return polarity indicating an increase or decrease in the brightness level of each pixel. Managing noise in these cameras is also different from conventional cameras because noise can easily cause incorrect events and increasing noise can disrupt camera performance [1]. Additionally, since the output of these cameras is different from that of conventional frame-based cameras, new methods need to be introduced for processing this output or ways need to be created to adapt them to the output of conventional cameras. One of the methods using event data is image reconstruction. Reconstructed images can be used in all normal algorithms, network architectures, and pre-trained weights without the need for adaptation. Nowadays, reconstructing image intensity using deep neural networks is common. Newer methods attempt to produce high-quality images in a shorter time, but each of these methods has its challenges. For instance, they may not work well when there are no events or the number of events is low. We expect to be able to preserve the parts of the scene that do not send events and update the parts that contain events. If we have few or no events for moments, some methods may not be able to reconstruct the image intensity as they do not consider the temporal relationship between frames and try to reconstruct an image, not a frame of video. Nevertheless, even methods that reconstruct the video and consider the relationship between frames also have difficulty in reconstructing these scenes. Another issue is the low quality of initial frame reconstruction in many methods, which requires initial time to reach acceptable quality levels.

In this article, we have used the E2VID model [2] as a basis due to its pioneering nature and high image quality. The E2VID model uses three encoder layers, each of which contains an LSTM. These LSTMs, which are responsible for transferring general scene information from the previous frame to the next one, do not always work well in some cases. In this approach, we attempted to modify the short-term and long-term memory update of the LSTM model and prevent image degradation. The advantage of our method is it can be applied to all LSTMs, and there is no need to retrain large networks such as E2VID. Improvement can be made by changing the updating method at test time.

979-8-3503-3015-1/23/$31.00 ©2023 IEEE

In summary, the contributions of the article are as follows:

1- Improving the quality of initial frames without the need for retraining and having very powerful processors.

2- Improving the quality of all frames by higher the contrast in images and expanding histograms.

3- Detecting times of event scarcity and changing the LSTM structure to improve reconstruction quality in these challenging conditions.

II. Related Work

Reconstructing image intensity from event streams is a challenging task. Initial approaches in this field considered assumptions about scene properties or camera motion. In terms of implementation, methods can be divided into two main categories. One method attempts to reconstruct images using optimization methods through estimating thresholds. Most of these methods use a base image to start with and apply events with calculated thresholds on the initial image. Although these methods are usually fast, their quality is still low. With the increasing use of deep neural networks, most methods now use these networks for reconstructing image intensity while require large data for training. Therefore, methods for simulating events from frames have been introduced [3]. Generally, the results of the first category have lower quality and more assumptions compared to the second category, but their frame reconstruction time is shorter. Brandli *et al.* [4] carried out one of the earliest works in this field in 2014. They estimated the intensity of each pixel by using a set of frames and events and a simple equation proportional to the difference between the number of positive and negative events. In 2016, Miyatani *et al.* [5] used a dictionary-based model for image reconstruction and noise reduction, which was based on learning. In 2020, Pan *et al.* [6] introduced an energy minimization-based approach that consists of two terms: one term constrains image brightness intensity to suppress noise and preserve edges, while the other one sharpens edges when selecting an appropriate threshold. In most neural network-based methods for reconstruction, a U-Net structure is used that is composed of an encoder-decoder. Additionally, most methods that aim to implement video frame sequences use a recursive model within their encoder to retain information from previous frames. However, some of the methods solve this problem by feeding previous frames into the network. In 2019, E2VID was introduced as a learning-based method for video reconstruction. In this method, a 3D tensor given to a U-Net recursive network include convLSTM in its encoder layer. Although this method has good quality, it takes much longer than HF and MR algorithms. The training data for this model is entirely simulation data but has performed well on real data. In 2020, Scheerlinck *et al.* [7] introduced a similar method that uses convGRU instead of convLSTM and is approximately three times faster than the previous method but has lower quality in challenging scenes compared to E2VID. Wang *et al.* [8] introduced another method that has been used deep neural networks for reconstruction in 2019. This method uses conditional generative networks or cGANs, to generate images. The strength of using cGANs is that there is no need to define a specific loss function and it can adapt its training loss to the data it was trained on. Research published in 2021 attempts to solve the problem of frame reconstruction in the initial moments. This article does not use a temporal loss function and controls the temporal relationship between frames with a layer called SPADE [9]. It places a conditional batch normalization layer for style transfer and manages the temporal relationship between frames. The next article generates images with higher spatial resolution. Wang *et al.* [10] reconstruct the images with low resolution and improves their quality, and then upsample them. This article consists of three phases: 1- Reconstruction, 2- Enhancement, and 3- High-resolution image generation. In each phase of this method, there are three networks: 1- Generator network that takes events as input and produces an image as output. 2-A network that takes the output image and tries to reconstruct the events from it. 3-A discriminator network. After that, Weng *et al.* [11] uses a hybrid CNN-Transformer approach to obtain local information and features from CNN and global features from Transformer. In this study, we have used the E2VID model as the base model, and by changing the structure of LSTM and the way cell and hidden are updated, we have improved both the quality of initial images and the overall quality of the video and prevented image degradation during times when there is no event.

III. Method

In this section, the structure of an LSTM is shown to explain how the cell and hidden are updated differently.

A. Overview

In Fig. 1, an LSTM cell is depicted. Its outputs are calculated as follows:

$$F_t = \sigma(U_f\, h_{t-1} + W_f\, x_t + B_f) \tag{1}$$

$$I_t = \sigma(U_i\, h_{t-1} + W_i\, x_t + B_i) \tag{2}$$

$$O_t = \sigma(U_o\, h_{t-1} + W_o\, x_t + B_o) \tag{3}$$

$$C\_G_t = tanh(U_{c\_g}\, h_{t-1} + W_{c\_g}\, x_t + B_{c\_g}) \tag{4}$$

$$C_t = F_t \times C_{t-1} + I_t \times C\_G_t \tag{5}$$

$$H_t = O_t \times tanh(C_t) \tag{6}$$

B. Enhancing the Quality of Initial Frames and Improving Quality of all frames

One of the reasons for poor quality in initial frames is the lack of contrast in images. However, this is not the only factor. At the first moment, the input of model does not have much difference with later moments but the emptiness of hidden and cell at the start causes low quality at the initial moments. This idea led us to find a way to place appropriate information inside hidden and cell by updating them several times with the same input.

When the outputs of LSTM are generated, instead of outputting them and passing them through the rest of the model, they are fed back into the model and the cell and hidden values for this LSTM are recalculated again. This solution has led to an improvement in the quality of initial frames. For this purpose, we have changed the structure of LSTM as follows and have tried to update a cell multiple

times instead of a single update or in some cases, consider an intermediate state for updating cell and hidden. In Fig. 2, an overview of our LSTM with two repetitions is shown. Initially, cell and hidden are initialized with zero values, and in the first update, the value of cell becomes equal to the information extracted from the input. At this stage, the value of cell has passed through tanH activation function and its maximum and minimum values are near -1 and 1. Then we repeat this operation once or several times. In the second repetition, the LSTM does not have new input (new information) and only hidden value has changed from zero. By looking at the maximum and minimum values of cell, we realize that one unit has been added to these values (-2 to 2). This increase shows that in this state, the model tries to reinforce itself at points where it previously had a value or information so that image contrast improves. After a while when cell takes a more stable state (since it holds long-term information), we no longer have significant changes in the minimum and maximum levels. In conclusion, this operation is completely different from just multiplying cell values by a specific number. Another aspect that has been investigated is using an average update for the cell, which in some cases resulted in better video quality. This update calculates the value by a weighted sum of the two previous cells and updates the hidden value with the new cell value. The method that improved the quality of the initial frames also improved the quality of the entire video and increase the contrast of the images. As it is evident from the results, initially networks should update more frequently (3-4 times), but after reaching a stable state, repeating once can improve contrast and details become more visible.

C. *Quality Improvement of Images in Absence or Lack of Events*

One of the issues with the E2VID method, is when there is no event, the image quality decreases gradually. To solve this problem, we change the LSTM, as before. The first point to note is that the input norm value has a direct relationship with the number of events and increases with them. When there is no input, the input norm is independent of the video we use for reconstruction and is approximately 64. Based on this point, we first identify when this problem is occurring and prevent updating the cell value in the first LSTM to keep acceptable

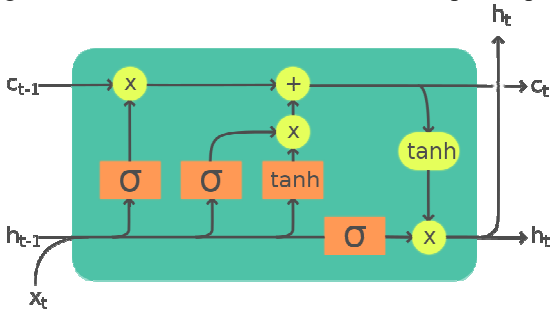

Figure 1. An LSTM cell

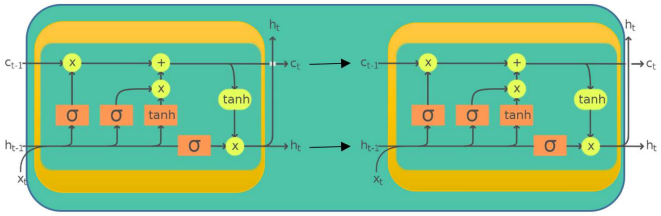

Figure 2: Our LSTM cell with two updates

quality for a longer period. However, since we need to examine the problem more generally at times when there are few events (not completely no events), we use a weighted average to update cell and hidden. First, we update LSTM values normally and if we name the updated cell as $cell_{new}$ and the updated hidden as $hidden_{new}$ then for the first LSTM output cell we have:

$$Cell = \frac{norm(input) - 64}{100} \times cell_{new} + \frac{164 - norm(input)}{100} \times Prev\_cell \quad (7)$$

Prev_cell and $Prev_{hidden}$ are input to the model from the previous step. If the input norm value is less than 164, we use Eq. (7) to update it; otherwise, it shows that there are enough events for image reconstruction so it updates normally. For updating hidden variables, we use Eq. (8):

$$hidden = \frac{norm(input) - 64}{100} \times hidden_{new} + \frac{164 - norm(input)}{100} \times Prev_{hidden} \quad (8)$$

Since the first LSTM provides input to subsequent LSTMs, preventing updates ensures that subsequent inputs are not damaged. On the other hand, updating the second and third LSTM causes their values to be up-to-date so if information returns to its initial state and we have enough events again as input, cell values do not need much time for initialization.

IV. Experiments and Results

Our base model is E2VID [2], and we used its pre-trained weights[2]. To evaluate our research, we used seven well-known datasets[3] that are publicly available. These datasets were captured with a DAVIS240C camera, and the image dimensions are 180x240. In these videos, the camera speed gradually increases, causing severe blurring in the reference images. For this reason, most evaluations are considered for the initial frames up to 550 frames. We compared our reconstructed images with ground-truth images from the datasets. To calculate error values in different metrics and to normalize the brightness intensity range, local histogram equalization was applied to all frames generated by the model and reference frames. In most evaluations, we used three metrics: SSIM (higher values are better) [12], MSE, and LPIPS [13]. Due to the different settings that the LPIPS metric can have, such as different versions of VGG models, different layers of VGG that can be selected, and different weights, it is not comparable with reported values in various articles. We

[2] Available at: http://rpg.ifi.uzh.ch/E2VID.html
[3] Available at: https://rpg.ifi.uzh.ch/datasets

compared our proposed method with two works: E2VID and SPADE. To compare with the SPADE method, we used results and images reported by the paper itself since it has not been implemented. The SPADE method was trained precisely on the dataset related to E2VID.

A. Results and Discussion

We compared our method (weighted average update) with the E2Vid method when the number of events is low. To do this, we manually cut off the event stream from frame 20 onwards and continued this cutoff until frame 300 (until a relative decrease in frame quality) and fed an empty tensor without events to the network. Finally, we calculated the average of three metrics for 280 frames without input on the sequences. In Table 1, we consider the last reconstructed frame by the model as the ground-truth frame when we have an input event. So the ground-truth frame is frame 20. As shown in the table, there is a significant improvement in results. This model has three LSTM units and we only updated the first LSTM using Eq. (7). Since LSTM units two and three are updated normally, with a new stream of events, the time of the initialization phase decreases (if a frame is destroyed, after restoring event stream, we need time to reach initial quality again like at the beginning of the video). As shown in Table 1, our model performed better in all metrics, with approximately 6% improvement in SSIM, 55% improvement in MSE, and 42% improvement in LPIPS. Fig. 3 also shows the output of both methods for frame 300 of the sequences. Table 2 indicates that after 300 frames if we feed the model with the event stream again, what the value of SSIM is. A higher SSIM value in our model indicates that it takes less time to achieve better quality after events occur.

Events may not always be completely cut off, and there may be moments or parts of the image where there are few events. To test our method and compare it with the baseline, we prepared a sequence. We first increased the number of frames for a part of the video using Super-SloMo [14] method with interpolation to have fewer events between each reference frame. Then the input events are normal until frame 20(with enough events) and from frame 20 onwards generated events with low events are entered into both our methods and the baseline. In our improved method, we had one more iteration inside each LSTM to improve quality until frame 20. Fig. 4 shows the output of both our models and E2VID for video slider_depth for frames from top to bottom: 0, 20, 150, 300, 400 and 500. In each image, the right side shows the reconstructed frame and the left side shows events related to that frame with red and blue colors. The input to our first LSTM had fewer events and the norm of input is less than 110 which was suitable for our test. Our method has better quality images due to two reasons: first by reducing the impact of LSTM updates and second due to the higher contrast present in images. Additionally, it was mentioned in the E2VID paper that parts of images with higher contrast tend to preserve their value over more frames.

Another issue under investigation was the quality of initial sequence are displayed. According to Fig. 5, the quality of the E2VID method is poor due to the emptiness of LSTM memory, and details are not visible. However, in the proposed method here, there is much more contrast in the image. Additionally, due to the way LSTM memories are updated, the quality of all frames is also higher.

Results for 10 initial frames based on average metrics are reported in Table 3. In this table, we use Re instead of repeat e.g. 2Re means that two stages of updating occurred within LSTM. Our method with four times repeat had an improvement of 11.9% in SSIM metric, 8.4% in MSE metric, and in LPIPS metric are almost the same.

Table 4 shows values for 10 initial frames mentioned in the SPADE article. Differences between our results and SPADE in E2VID Method are due to different post-processing applied to output images.

Histograms for five specific frames for both E2VID and our improved method with two times repeat are shown in Fig. 6. To evaluate the contrast, we used the RMS metric that calculates standard deviation of each pixel in the frame, and at the end, we average these values and report this for RMS of sequences. Table 5 shows RMS values for the first 200 frames (except Slider_Depth which has only 87 frames). A higher value means higher image contrast. Our method has a significantly higher RMS value than E2VID. The value reported in the SPADE paper for their method is an average of

TABLE 1. Results of our method for lake of events. Our method has better result in all three metrics.

Dataset	MSE		SSIM		LPIPS	
	E2VID	***Ours***	***E2VID***	***Ours***	***E2VID***	***Ours***
Shapes_6dof	0.0171	**0.0064**	0.9303	**0.963**	0.1081	**0.0587**
Dynamic_6dof	0.0086	**0.0027**	0.8426	**0.9309**	0.2184	**0.0484**
Calibration	0.0105	**0.0074**	0.8309	**0.8844**	0.1846	**0.0672**
Boxes_6dof	0.063	**0.0249**	0.686	**0.7403**	0.1802	**0.1566**
Poster_6dof	0.0446	**0.0234**	0.6364	**0.6806**	0.1852	**0.1717**
Mean	0.0287	**0.0129**	0.7852	**0.8398**	0.1753	**0.1005**

TABLE 2. Comparison of our method with E2VID in SSIM when we do not have any event for a few seconds.

	SSIM	
Data Sequence	***E2VID***	***Ours***
Shapes_6dof	0.3799	**0.3857**
Dynamic_6dof	0.2276	**0.2723**
Calibration	0.2873	**0.3808**
Boxes_6dof	0.3498	**0.4491**
Poster_6dof	0.3602	**0.4026**
Slider_depth	0.3525	**0.3881**
Office_zigzag	0.2826	**0.3314**

47.4627, while they reported a value of 38.2259 for E2VID. So SPADE is almost 24% better than E2VID and our method is about 31% better than E2VID. This difference in values for E2VID may be due to differences in the number of calculated frames or other post-processing steps.

Due to fast movements in some sequences and the fact that with fast movement, the reference image also becomes blurry and low quality, we calculated metrics for the first 550 frames of the videos to compare our method. Table 6 shows the quality of videos for the first 550 frames. As shown in Table 6, our method with one additional iteration inside LSTM improved SSIM metric by about 2%, MSE metric by 7%, and LPIPS metric by 2%. The value reported in the SPADE paper for the first 550 frames, shows similar numbers for both methods (E2VID and SPADE) and they do not have an improvement in metrics. In Fig. 7, you can see an example of our model output, E2VID, and SPADE.

It should be noted that our model increases frame reconstruction time due to changes in update and cycle repetition within LSTM, taking about 0.014 seconds on our hardware for the E2VID model and about 0.02 seconds with one iteration for our model to generate output, with an approximate increase of 0.006 seconds per additional LSTM iteration.

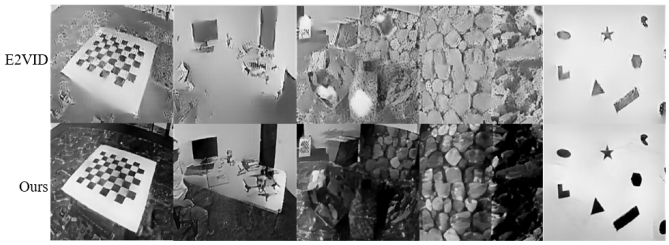

Figure 3: Frame 300, frame degradation in the absence of events.

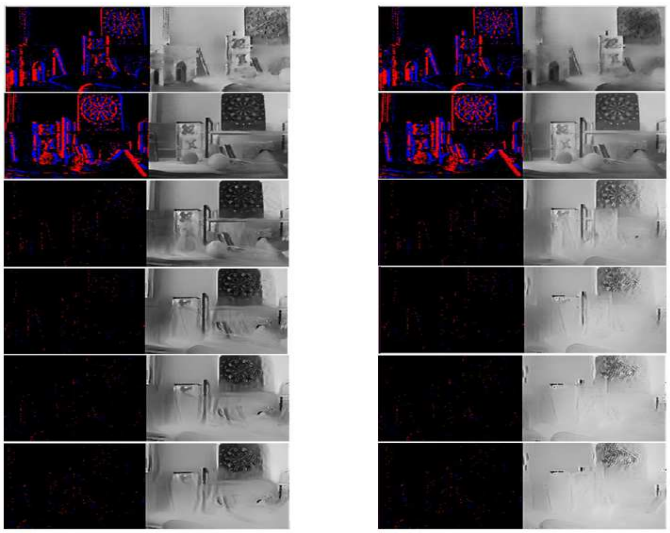

Figure 4: The right side shows the E2VID output and the left side shows the model's output for not enough events from top to bottom frame: 0, 500, 400, 300, and 150. In each of the images, the right side shows the visualization of the reconstructed frame by the model, and the left side shows the corresponding events for that frame.

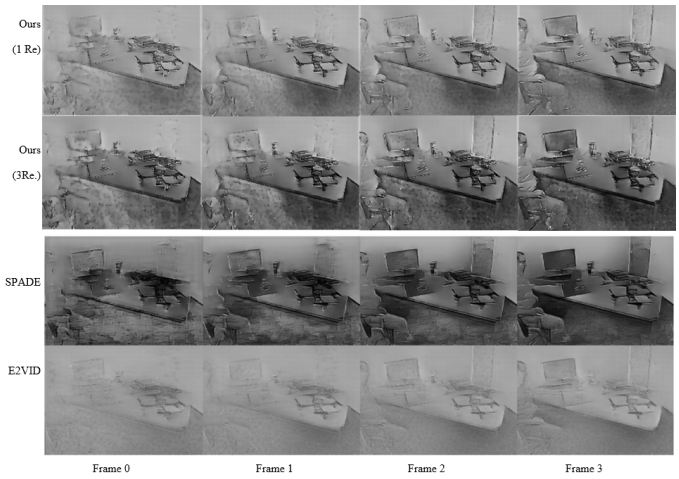

Figure 5: Four initial frames of the Dynamic_6dof sequence

TABLE 3: Comparison to E2VID for the first 10 initial frames. Our method improves an average of 11.9% in SSIM.

Dataset	MSE			SSIM			LPIPS		
	E2VID	*Our (4Re)*	*Our (2Re)*	*E2VID*	*Our (4Re)*	*Our (2Re)*	*E2VID*	*Our (4Re)*	*Our (2Re)*
Shapes_6dof	**0.1699**	0.2034	0.2022	0.1838	**0.2373**	0.1934	0.5666	0.5614	0.5608
Dynamic_6dof	0.1398	0.1971	0.1783	0.2171	0.2115	0.2204	0.4914	0.5418	0.5234
Calibration	0.1504	0.1354	**0.1236**	0.182	0.1914	0.2104	0.5468	0.5304	**0.5249**
Office	0.1806	0.2034	**0.1463**	0.1231	**0.1629**	0.154	0.5348	0.5064	**0.502**
Boxes_6dof	0.1322	0.1322	0.1263	0.1802	0.2178	0.2231	0.5314	0.5082	**0.4976**
Slider	0.0667	0.0611	**0.0598**	**0.359**	0.3451	0.3545	**0.4501**	0.4779	0.4557
Poster_6dof	0.15	0.1384	**0.1358**	0.2291	**0.2793**	0.2729	0.4836	0.4895	**0.4712**
Mean	0.1414	0.153	0.1389	0.2106	**0.235**	0.2327	0.515	0.5165	0.5051

TABLE 4: Reported value by SPADE for 10 initial frames [9]

	MSE		SSIM	
	E2VID	*SPADE*	*E2VID*	*SPADE*
Mean	0.2003	**0.1685**	**0.2381**	0.248

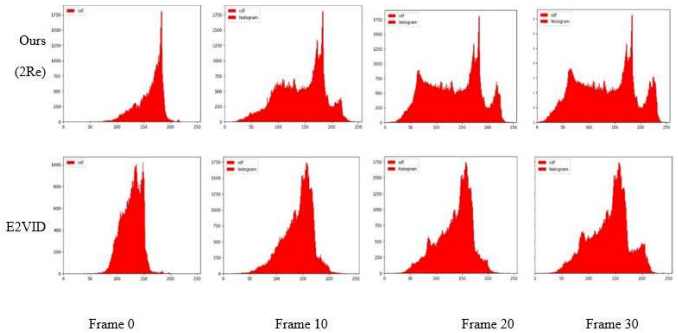

Figure 6: Comparison between the histogram of our reconstructed image and the E2VID image. Our reconstruction shows an extended histogram and an extended histogram means high contrast.

V. CONCLUSION

The method presented in this study does not require retraining and by changing the updates of LSTM during test time, we improve the results. In addition to the improvement in evaluation metrics, this model also yields very desirable results according to the RMS metric, which measures contrast in images. The proposed method performs better than the E2VID method with an 11.9% improvement in the first 10 frames and a 2% improvement in entire videos in SSIM metric.

Finally, this method is proposed to manage when there are few or no existing events, which performs 6% better than E2VID for 280 frames according to the comparison.

TABLE 5. Experiment on RMS contrast metric. Results show improvement in image contrast

	RSM contrast	
Data Sequence	***E2VID***	***Ours (1Re)***
Shapes_6dof	31.7713	**39.59**
Dynamic_6dof	52.6444	**70.5063**
Calibration	54.4538	**74.0012**
Office_zigzag	41.9373	**61.33**
Boxes_6dof	41.0182	**49.0034**
Slider_depth	40.659	**49.5074**
Poster_6dof	42.53	**57.7801**
Mean	43.5734	**57.3883**

TABLE 6. Comparison to E2VID for the first 550 frames. our method outperforms E2VID with an average of 2% in SSIM and 7% in MSE.

Dataset	**MSE**		**SSIM**		**LPIPS**	
	E2VID	***Ours***	***E2VID***	***Ours***	***E2VID***	***Ours***
Shapes_6dof	0.1879	**0.1266**	0.2774	**0.2969**	0.5324	**0.5073**
Dynamic_6dof	**0.0714**	0.0746	0.3848	**0.4507**	0.3886	**0.3781**
Calibration	0.0416	**0.0385**	**0.4191**	0.4171	0.4219	**0.4185**
Office_zigzag	**0.0685**	0.0718	0.238	**0.2432**	**0.418**	0.4241
Boxes_6dof	**0.0348**	0.0442	**0.5533**	0.5321	0.36	**0.3521**
Slider_depth	0.0818	**0.0716**	0.4668	**0.4834**	**0.3928**	**0.3928**
Poster_6dof	**0.042**	0.06	**0.5754**	0.5443	0.327	**0.3199**
Mean	0.0754	**0.0696**	0.4164	**0.4239**	0.4058	**0.3989**

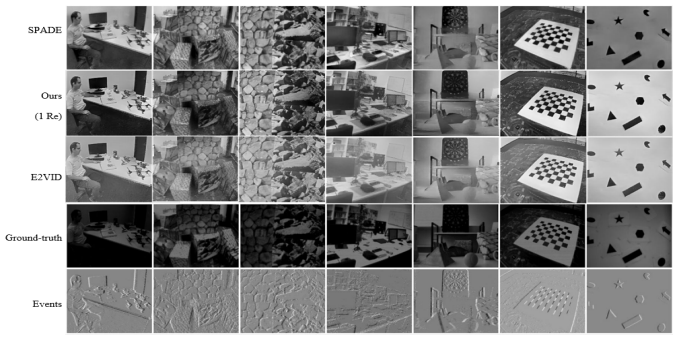

Figure 7: Comparison of our model output with E2VID and SPADE, our method can reconstruct an image with low overhead and better quality

REFERENCE

[1] G. Gallego, D. Tobi, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis and D. Scaramuzza, "Event-based Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2020.

[2] H. Rebecq, R. Ranftl, V. Koltun, D. Scaramuzza, "High Speed and High Dynamic Range Video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2019.

[3] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió and D. Scaramuzza, "Video to Events: Recycling Video Datasets for Event Cameras," *The Conf. on Comput. Vis. and Pattern Recog. (CVPR),* 2020.

[4] C. Brandli, L. Muller and T. Delbruck, "Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor," *IEEE Int. Symp. Circuits Syst. (ISCAS),* 2014.

[5] Y. Miyatani, S. Barua and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," *IEEE Winter Conf. Appl. Comput. Vis. (WACV),* 2016.

[6] L. Pan, R. Hartley, C. Scheerlinck, M. Liu, X. Yu and Y. Dai, "High Frame Rate Video Reconstruction based on an Event Camera," *arXiv e-prints*, 2019.

[7] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. E. Mahony and D. Scaramuzza, "Fast Image Reconstruction with an Event Camera," *IEEE Winter Conf. on Applications of Comput. Vis.* ,2020.

[8] L. Wang, M. Mostafavi, Y.-S. Ho and . K.-J. Yoon, "Event-based High Dynamic Range Image and Very High Frame Rate Video Generation using Conditional Generative Adversarial Networks," *In IEEE Conf. Comput. Vis. Pattern Recog(CVPR),* 2019.

[9] P. Rodrigo, G. Cadena, Y. Qian, M. Yang and C. Wang, "SPADE-E2VID: Spatially-Adaptive Denormalization for Event-Based Video Reconstruction," *IEEE Transactions on Image Processing,* 2021.

[10] L. Wang, T.-K. Kim and K.-J. Yoon, "EventSR: From Asynchronous Events to Image Reconstruction, Restoration, and Super-Resolution via End-to-End Adversarial Learning," *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] W. Weng, Y. Zhang and Z. Xiong, "Event-based Video Reconstruction Using Transformer," *International Conf. on Comput. Vis.(ICCV),* 2021.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, Apr. 2004.

[13] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018.

[14] H. Jiang, D. Sun, V. Jampani, MH. Yang, E. Learned-Miller, J. Kautz. "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation". *In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018.