

# Spk2VidNet: A Hierarchical Recurrent Architecture for High-Fidelity Video Reconstruction from Long Spike-Camera Streams

Yuanlin Wang<sup>1</sup>, Ruiqin Xiong<sup>1\*</sup>, Jiyu Xie<sup>2</sup>, Zhenkun Zhu<sup>1</sup>, Zhaofer Yu<sup>1</sup>, Xiaopeng Fan<sup>3</sup>, Tiejun Huang<sup>1</sup>

<sup>1</sup>State Key Laboratory of Multimedia Information Processing,  
School of Computer Science, Peking University

<sup>2</sup>Shanghai Radio Equipment Research Institute

<sup>3</sup>School of Computer Science and Technology, Harbin Institute of Technology

{wangyuanlin, zkzhu}@stu.pku.edu.cn

{rqxiong, yuzf12, tjhuang}@pku.edu.cn    xjy646@mail.ustc.edu.cn    fxp@hit.edu.cn

## 8. Experimental Details

### 8.1. Inference Details

During evaluation, the length of spike sequences that can be processed in a single inference is constrained by GPU memory. For very long spike sequences, we adopt the same strategy as in training. Specifically, we divide the input sequence into multiple smaller segments and perform inference on each segment sequentially (refer to Fig. 6).

For segments from the same sequence, the last states of the  $d$ -th segment are stored in the feature buffer and used for the inference of the  $(d + 1)$ -th segment. The key difference from the training phase is that, during inference, the states in the buffer do not undergo the detach operation, as no gradient backpropagation is involved.

### 8.2. Loss Function

We use  $\mathcal{L}_1$  loss to train the model. The loss function is formulated as follows:

$$\mathcal{L}_1 = \frac{1}{N} \sum_{i=0}^{N-1} \|\mathcal{I}_i / \eta - \mathcal{I}_i^{gt}\|_1,$$

where  $\mathcal{I}_i$  is the generated high resolution (HR) image at time  $t_i$ ,  $\eta$  is the photoelectric conversion rate,  $\mathcal{I}_i^{gt}$  is the ground truth.

### 8.3. Computational Complexity (MACs)

In addition to the model parameters, runtime, and GPU memory cost provided in Tab. 1, as shown in Tab. 3, we compare the proposed Spk2VidNet with other spike camera super resolution (SCSR) methods in terms of computational complexity (MACs, multiply-accumulate operations). The MACs in Tab. 3 represent the computational cost required

Table 3. Comparisons on MACs for  $\times 4$  SCSR models.

Method	MACs/G
VidarSR [5]	16537.4
SpikeSR-Net [6]	7273.2
Spk2SRImgNet [4]	1238.5
SCSRNet [3]	1057.8
Spk2VidNet (Ours)	<b>298.1</b>

to infer a single HR image, tested on low resolution (LR) input of size  $180 \times 320 \times 101$  with  $\times 4$  SCSR models.

### 8.4. More Quantitative Analysis

In addition to the quantitative results presented in Tab. 1 of the main paper, we present PSNR curves on several sample sequences randomly selected from the REDS-LSSR and Adobe240-LSSR evaluation datasets ( $\times 4$ ), as shown in Figs. 8 and 9. The size of each sample in the datasets is listed in Tab. 4.

The frame index in Figs. 8 and 9 corresponds to the frame index in the spike sequence. As mentioned in Sec. 5.3, since the compared methods cannot reconstruct 4 initial and 4 final boundary frames for a spike sequence, we consider only the central portion of the ground-truth image sequence for each sample. Specifically, for the REDS-LSSR dataset, we compare the reconstruction results on the middle 37 frames of each sequence ( $45 - 4 \times 2 = 37$ ). For the Adobe240-LSSR dataset, we compare the reconstruction results on the middle 82 frames of each sequence ( $90 - 4 \times 2 = 82$ ). These curves offer a more detailed view of the SR reconstruction performance over time and further highlight the improvements achieved by the proposed Spk2VidNet.

\*Corresponding author.

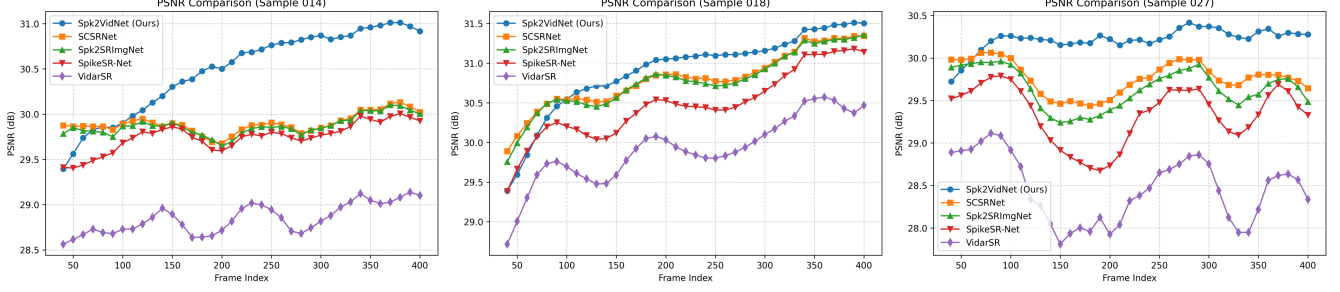


Figure 8. PSNR comparison on samples from the REDS-LSSR evaluation dataset.

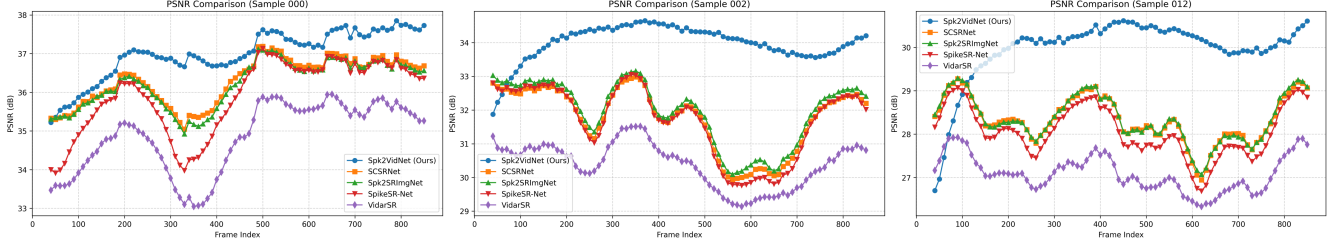


Figure 9. PSNR comparison on samples from the Adobe240-LSSR evaluation dataset.

Table 4. The size of each sample in REDS-LSSR and Adobe240-LSSR datasets. The size is represented in the format of height  $\times$  width  $\times$  temporal length.

Datasets	SR scale	Spike stream	Ground truth images
REDS-LSSR	$\times 4$	$180 \times 320 \times 461$	$720 \times 1280 \times 45$
	$\times 8$	$90 \times 160 \times 461$	$720 \times 1280 \times 45$
Adobe240-LSSR	$\times 4$	$180 \times 320 \times 911$	$720 \times 1280 \times 90$
	$\times 8$	$90 \times 160 \times 911$	$720 \times 1280 \times 90$

## 8.5. More Visual Results

To better understand the spike camera super-resolution (SCSR) task, we provide a *video* showcasing continuous SR reconstruction results on real-captured spike data. The real spike data are recorded at 20,000 Hz, and our method reconstructs HR frames at the same temporal resolution, i.e., 20,000 Hz. For presentation purposes, the reconstructed frames are played at 30 Hz (about  $667\times$  slowdown) to clearly present SR results. Consequently, a 5-second sequence in the video corresponds to only 7.5 milliseconds of actual motion.

Additionally, we provide more visual comparisons on both real-captured and synthetic spike data to further demonstrate the superior performance of Spk2VidNet, as shown in Figs. 12 to 14. Compared to other methods, Spk2VidNet produces HR images with finer details and better visual quality. Please enlarge the figure for better comparison.

## 9. More Ablation Study

### 9.1. Multi-Frame Consistency Guided Alignment

Within propagation block, Spk2VidNet utilizes multi-frame consistency guided alignment (MFCA) module to align neighboring features with current feature. To investigate the module design in MFCA, we further perform ablation studies, as shown in Tab. 5 (“w.o.” stands for “without”).

In case A, we remove MFCA module from Spk2VidNet, forming the baseline without any alignment. In case B, we replace MFCA with independent alignment (INAlign), where the same alignment module is used to separately align each neighboring feature to the current feature, without sharing motion cues or mutually refining. Specifically, as shown in Fig. 10, deformable convolution [1, 2] is employed to independently align  $F_{i-2}$  and  $F_{i-1}$  to  $Y_i$ . In case C, we remove mutual refinement in MFCA, leaving only joint motion estimation among neighboring features and current feature. Comparisons among cases A-D demonstrate the effectiveness of joint motion estimation and mutual refinement design in MFCA.

### 9.2. Content-Aware Modulation Fusion

After aligning neighboring features with current feature using MFCA module, Spk2VidNet adaptively fuses the aligned features with current one using content-aware modulation fusion (CMF) module, leveraging temporal correlations to suppress spike fluctuations.

To investigate the detailed design in CMF module, we consider two ablation cases, as shown in Tab. 6. In case (A), we remove CMF module from Spk2VidNet, forming

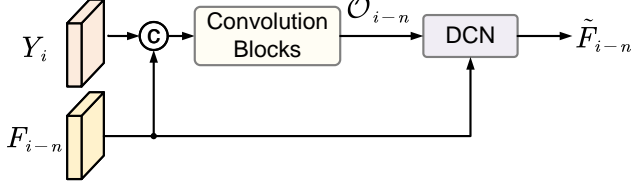


Figure 10. Illustration of independent alignment for each neighboring feature  $F_{i-n}$ .

Table 5. Ablation study of MFCA module on REDS-LSSR evaluation dataset ( $\times 4$ ).

Case	Setting Description	PSNR $\uparrow$	SSIM $\uparrow$
A	Removing MFCA from Spk2VidNet	29.27	0.8273
B	Replace MFCA with INAlign	29.67	0.8385
C	w/o. mutual refinement in MFCA	29.76	0.8421
D	The final model	<b>29.79</b>	<b>0.8432</b>

Table 6. Ablation study of CMF module on REDS-LSSR evaluation dataset ( $\times 4$ ).

Case	Setting Description	PSNR $\uparrow$	SSIM $\uparrow$
(A)	Removing CMF from Spk2VidNet	29.41	0.8325
(B)	w/o. MDM in CMF	29.70	0.8408
(C)	The final model	<b>29.79</b>	<b>0.8432</b>

the baseline. In case (B), we remove the multi-dilation modulation (MDM) submodule in CMF. In this setup, the aligned features and current feature are directly concatenated and passed through a convolution block followed by three residual blocks. The comparison between cases (B) and (C) demonstrates the effectiveness of the MDM submodule in CMF.

### 9.3. The Number of Propagation Features $K$

Value  $K$  denotes the number of preceding features used to enhance the current feature, and we explore the impact of  $K$  on SCSR performance. Based on the results in Tab. 7, we select  $K = 2$  for the final model.

Table 7. Ablation study on  $K$  values.

$K$	PSNR $\uparrow$	SSIM $\uparrow$	Params(M)	Runtime(ms)
1	29.65	0.8394	2.01	27
2	29.79	0.8432	3.73	43
3	29.82	0.8446	4.47	53

### 9.4. Propagation Layers

We perform ablation study on propagation layers, as presented in Tab. 8. The results show that three-level prop-

agation brings marginal gains with higher cost, while two-level propagation achieves the best quality–efficiency trade-off. Therefore, we adopt two-layer propagation design in Spk2VidNet.

Table 8. Ablation study on propagation layers.

Propagation Layer	PSNR $\uparrow$	SSIM $\uparrow$	Params(M)	Runtime(ms)
1	29.66	0.8383	2.30	27
2	29.79	0.8432	3.73	43
3	29.82	0.8428	5.16	57

### 9.5. Training Strategy

As mentioned in Sec. 6, we evaluate the impact of segment-wise training with state transfer. When each segment is trained independently, the SR reconstruction achieves 29.58 dB / 0.8377 PSNR / SSIM, which is lower than the performance achieved using the proposed training strategy (29.79 dB / 0.8432). In addition to the quantitative results, we further present PSNR curves for sample sequences from the REDS-LSSR and Adobe240-LSSR evaluation datasets, as shown in Fig. 11. These curves indicate that the overall PSNR trends under the two training strategies are similar. While the SR performance on the initial frames is comparable, Spk2VidNet with segment-wise training and state transfer consistently outperforms independent training as the sequence progresses. This comparison further demonstrates the effectiveness of the proposed segment-wise training with state transfer strategy.

### References

- [1] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [2] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1954–1963, 2019. 2
- [3] Yuanlin Wang, Yiyang Zhang, Ruiqin Xiong, Jian Zhang, Xinfeng Zhang, and Tiejun Huang. Super-resolving dynamic scenes with spike camera via multi-frame sequential alignment with motion propagation. *IEEE Transactions on Image Processing*, 34:6537–6549, 2025. 1
- [4] Yuanlin Wang, Yiyang Zhang, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, and Tiejun Huang. Spk2SRImgNet: Super-resolve dynamic scene from spike stream via motion aligned collaborative filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11416–11426, 2025. 1
- [5] Xijie Xiang, Lin Zhu, Jianing Li, Yixuan Wang, Tiejun Huang, and Yonghong Tian. Learning super-resolution reconstruction for high temporal resolution spike stream. *IEEE*

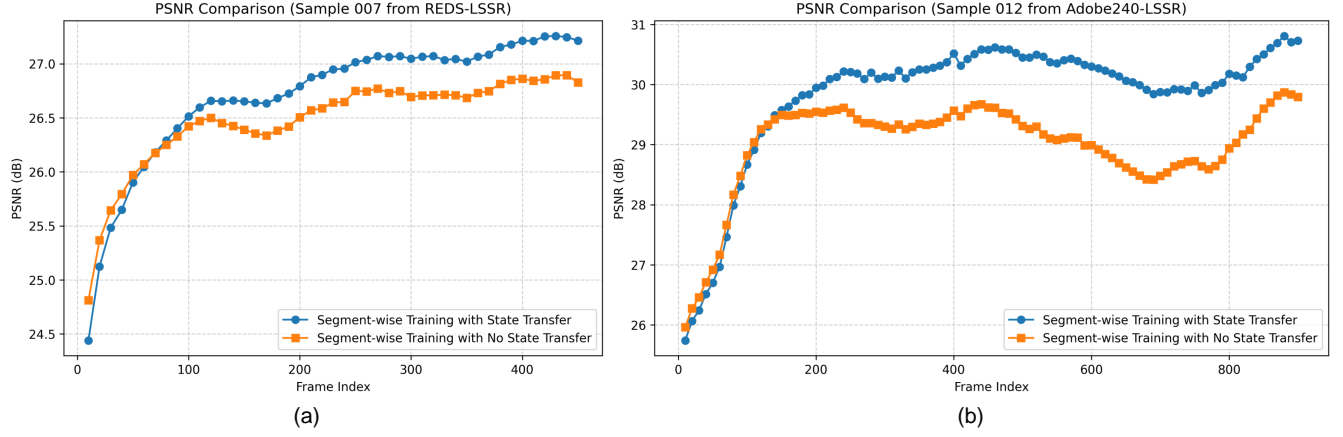


Figure 11. PSNR comparison of Spk2VidNet with different training strategies. (a) A sample from REDS-LSSR evaluation dataset. (b) A sample from Adobe240-LSSR evaluation dataset.

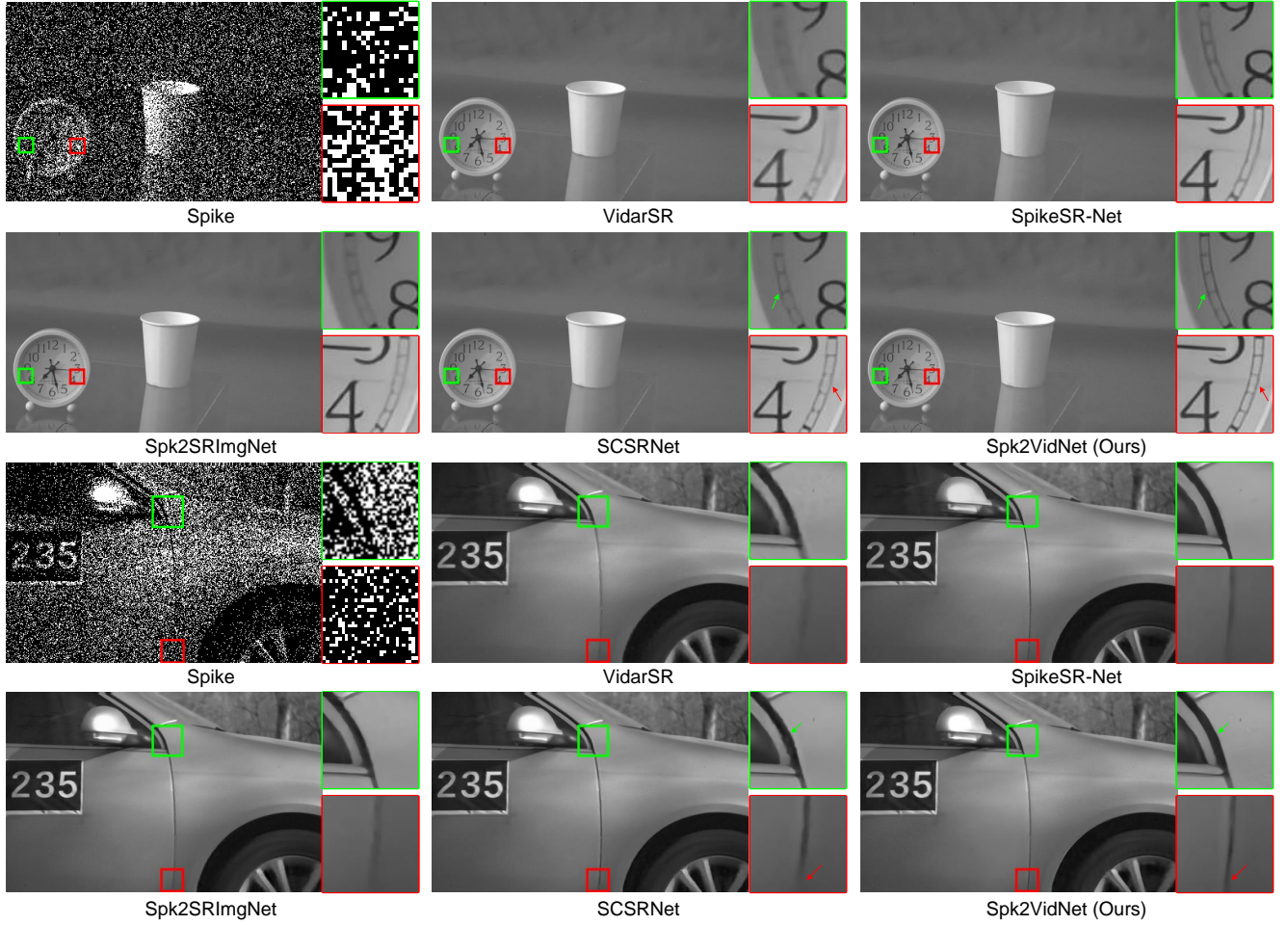


Figure 12. Visual results ( $\times 4$ ) on real-captured spike data. The first scene is an indoor scene recorded by a high-speed moving spike camera. The second scene is a running car. Please enlarge the figure for better comparison.



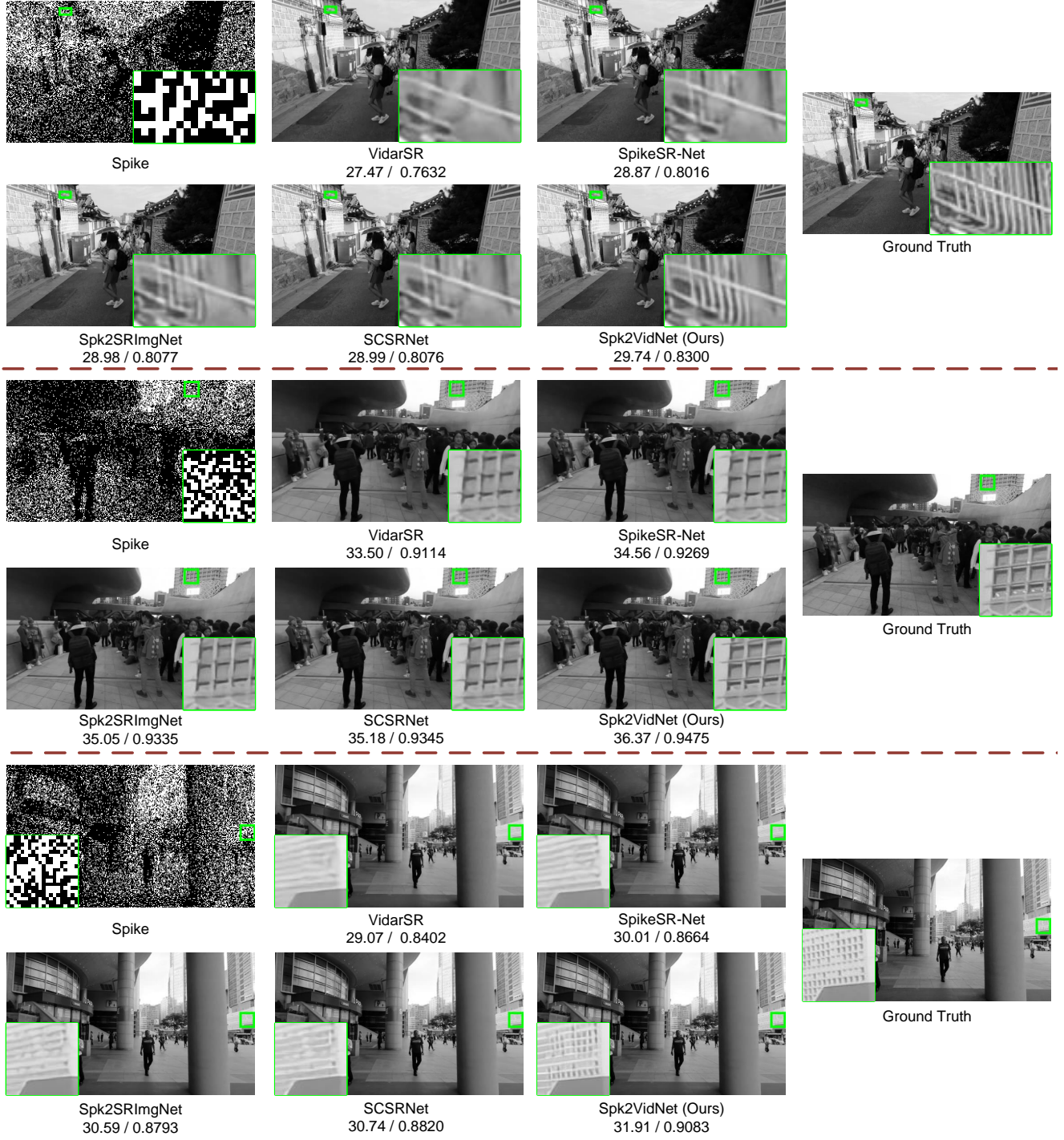


Figure 13. Visual comparison ( $\times 4$ ) on the REDS-LSSR data. The values below each image represent ‘PSNR / SSIM’ metrics. Please enlarge the figure for better comparison.

for neuromorphic spike camera. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3579–3587, 2023.

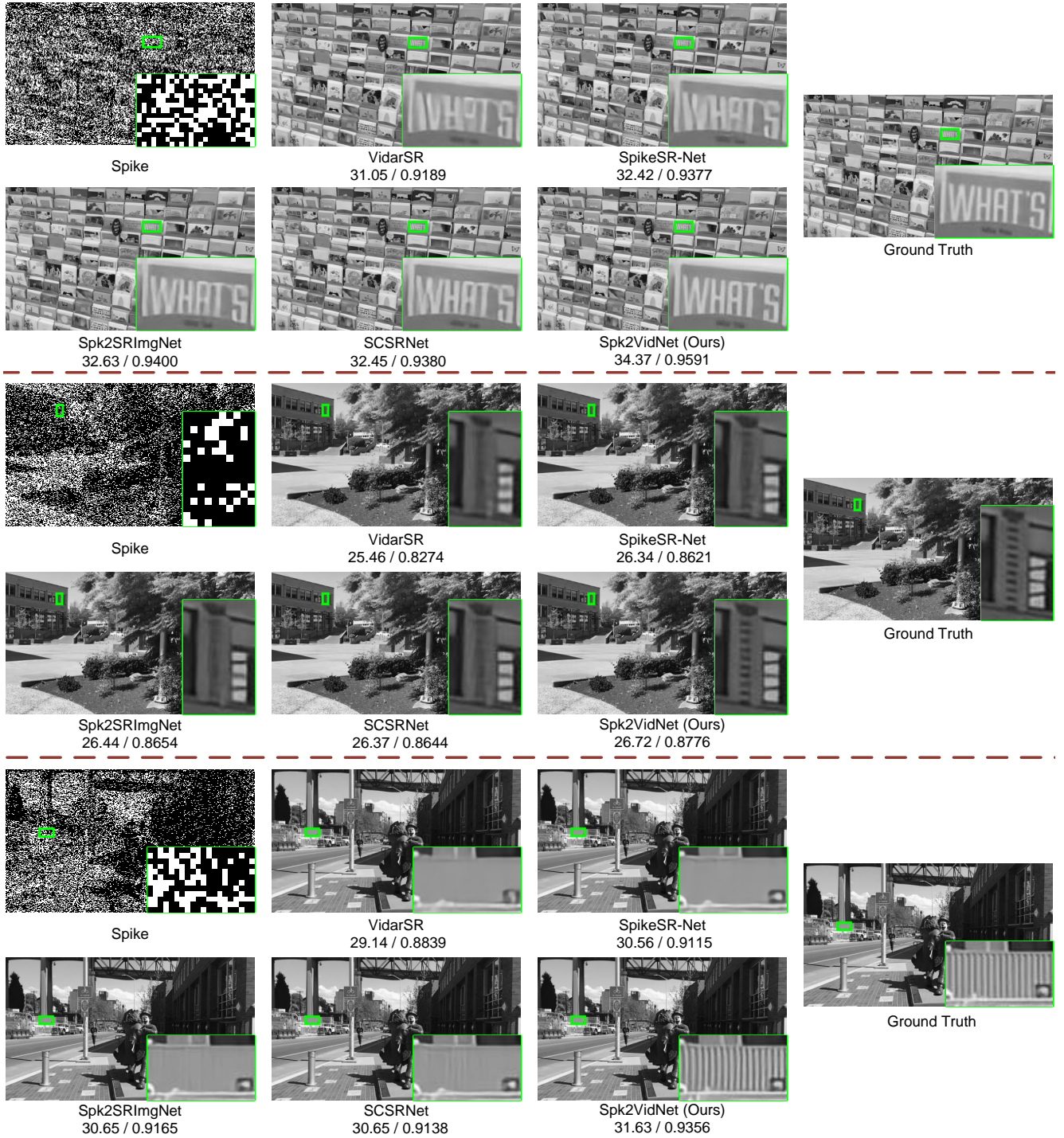


Figure 14. Visual comparison ( $\times 4$ ) on the Adobe240-LSSR data. The values below each image represent 'PSNR / SSIM' metrics. Please enlarge the figure for better comparison.