Supplementary Material – Synthetic Aperture Imaging with Events and Frames

Wei Liao[1]*, Xiang Zhang[1]*, Lei Yu[1]†, Shijie Lin[2], Wen Yang[1]†, Ning Qiao[3]
[1]Wuhan University, Wuhan, China
[2]The University of Hong Kong, Hong Kong, China
[3]Chengdu SynSense Tech. Co. Ltd., Chengdu, China
{weiliao,xiangz,ly.wd,yangwen}@whu.edu.cn,lsj2048@connect.hku.hk,ning.qiao@synsense.ai

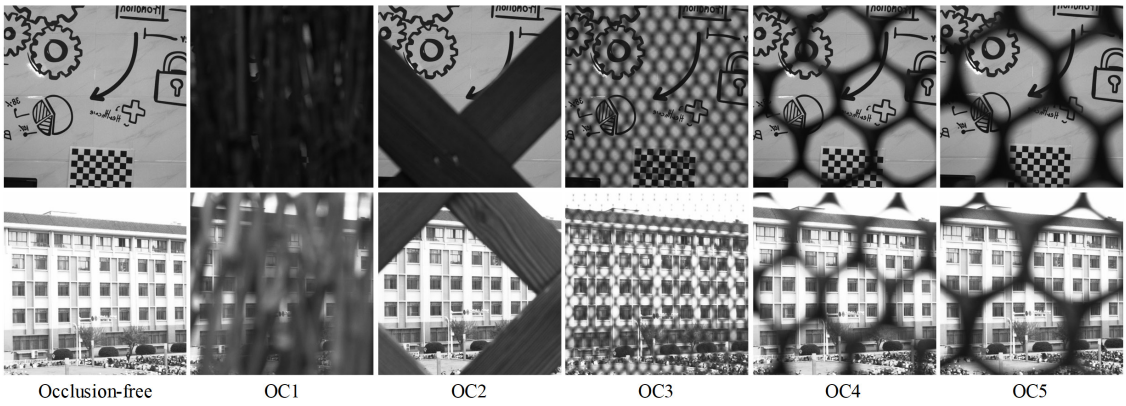

Figure 1. Five types of occlusions composed of extremely dense occlusion (OC1) and sparse occlusions (OC2-5) in EF-SAI dataset. The first and second rows correspond to camera views of indoor and outdoor scenes under different types of occlusions.

1. The EF-SAI Dataset

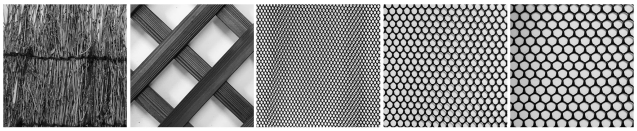

Figure 2. The samples of occlusions made of random thron fences (OC1), regular wooden grids (OC2) and regular plastic grids (OC3-5).

Table 1. Detailed data distribution of the proposed EF-SAI dataset.

Type	OC1	OC2	OC3	OC4	OC5	Summary
Indoor	245	100	150	150	150	795
Outdoor	148	0	13	14	18	193
Total	393	100	163	164	168	988

Fig. 1 and Fig. 2 shows the five types of occlusions composed of extremely dense occlusions (OC1) and sparse occlusions (OC2-5) in EF-SAI dataset. Specifically, the OC1 is made of random thorn fences, and the OC2 is made of regular wooden grids which enables the camera to observe larger fields of targets compared to OC1. The OC3-5 are made of regular plastic grids with different grid sizes. Tab. 1 summaries the number of data pairs in indoor and outdoor scenes for each type of occlusion.

2. Network Architectures

The detailed architecture of EF-SAI-Net is presented in Fig. 3, the height H and width W of the network inputs and outputs are set to (256,256). Let CnB(R)s-k denotes an $n \times n$ Convolution-BatchNorm (-ReLU) layer with k kernels and stride s. ResB-k denotes a residual block composed of a C3BR1-k layer and a C3B1-k layer with k kernels. DeCnB(R)s-k denotes an $n \times n$ fractional-strided-Convolution-BatchNorm (-ReLU) layer with k kernels and stride s. Then, the CNN-based sub-encoder in MF encoder consists of three convolutional layers: C3-8, C5-16, C7-3 and skip connections. The SNN-based sub-encoder in MF encoder consists of three spiking layers: S1-8, S3-16, S5-32 and skip connections. The CNN-based MF decoder can be described as C7BR1- 64, C3BR2-128, C3BR2-256, 9 ResB-256 blocks, DeCBR2-128, deCBR2-64 and a C7T-1 layer, where C7T-1 denotes a 7 $\times$ 7 Convolution-Tanh layer with 1 kernel.

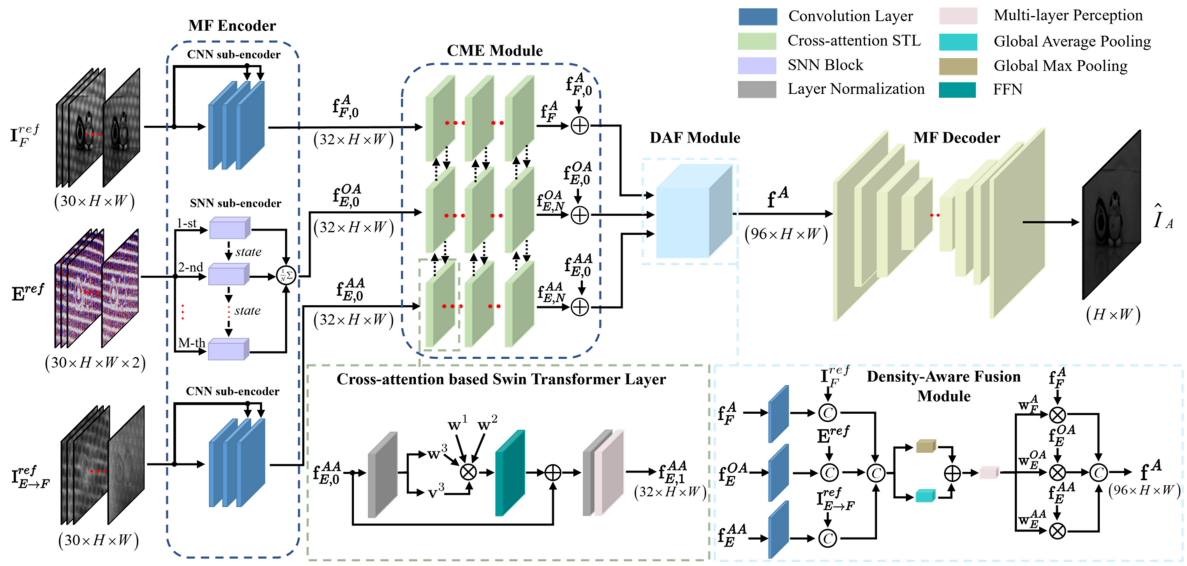

Figure 3. Detailed architecture of proposed EF-SAI-Net. The network is based on an Encoder-Decoder architecture, the input multi-modal signals are firstly fed into a Multi-modal Feature (MF) Encoder composed CNN- and SNN-based sub-encoders for feature extraction. Then, a Cross-Modal Enhancement (CME) module and a Density-Aware Fusion (DAF) module are used to mutually enhance and adaptively fuse the multi-modal features. Finally, a CNN-based MF Decoder is employed to reconstruct the occlusion-free visual images.

(a) Indoor scenes (b) Outdoor scenes

Figure 4. Occluded views of indoor (left four) and outdoor (right four) scenes under occlusions with different densities.

3. Extra Experimental Results

3.1. Extra Qualitative Comparisons

Fig. 5 and Fig. 6 show the extra qualitative comparisons of reconstructed images with error maps for indoor and outdoor scenes, which correspond to the occluded scenes in Fig. 4. For frame-based methods, the results of F-SAI+ACC are severely blurred since it equally treats the light information from targets and occlusions. The learning-based F-SAI+CNN effectively improves the quality of reconstructed images but is still disturbed by occlusions due to the limited number of observations. The performance of F-SAI+Inpainting significantly drops when occlusion information becomes dominant and is also limited by the inconsistent inpainted regions on different frames. For event-based methods, the E-SAI+ACC suffers from strong event noises and thus generates low-quality images. The E-SAI+Hybrid effectively suppresses the noise by employing SNN-based encoders. However, E-SAI+Hybrid can hardly reconstruct image details under sparse occlusions since events often respond to high-contrast areas and ignore low-contrast textures. The proposed EF-SAI-Net marries the merits of events and frames and employs the CME and DAF modules to adaptively fuse light information from multi-modal signals, achieving the best and consistent performance under different densities of occlusions.

3.2. Extra Qualitative Ablation Studies

Fig. 7 illustrates the qualitative comparisons of ablation results. From the reconstructed images and error maps, both CME module and DAF module are important for disturbance suppression and cross-modal signal fusion. The I^{ref} brings sufficient information of targets under sparse occlusion because the traditional camera directly captures noise-free intensity values of observed scene. The E^{ref} can provide a lot of available information under dense occlusions due to the continuous observation of the event camera. The $\mathrm{I}^{\mathrm{ref}}_{\mathrm{E}\to\mathrm{F}}$ is also helpful to improve the performance of SAI. In summary, the proposed EF-SAI-Net takes both events and frames as inputs and uses the CME and DAF modules to efficiently extract information to reconstruct occluded targets under a variety of occlusions.

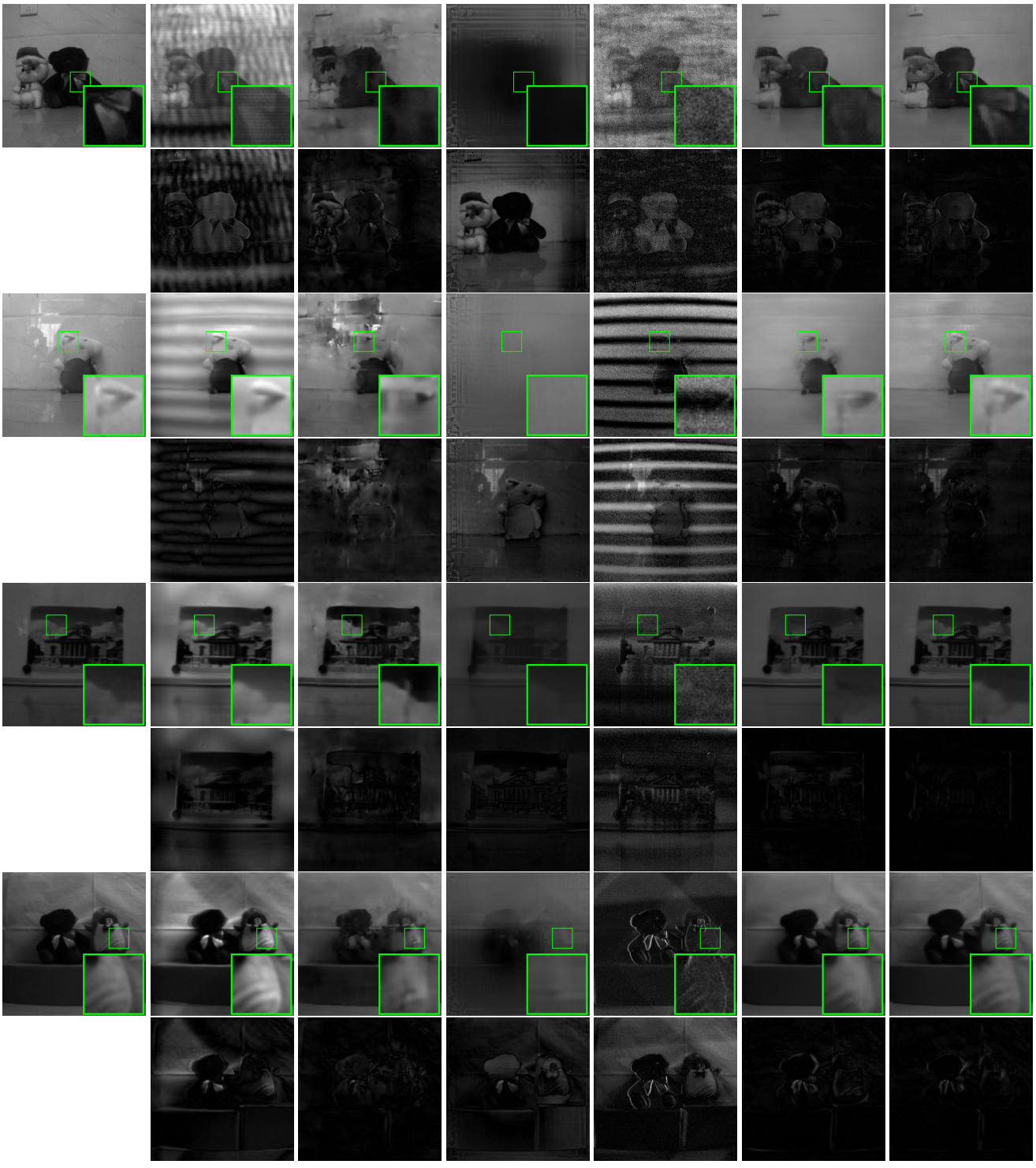

Figure 5. Qualitative comparisons of reconstructed results with corresponding error maps on the indoor scenes under dense occlusions (1st to 2nd rows) and sparse occlusions (3rd to 8th rows).

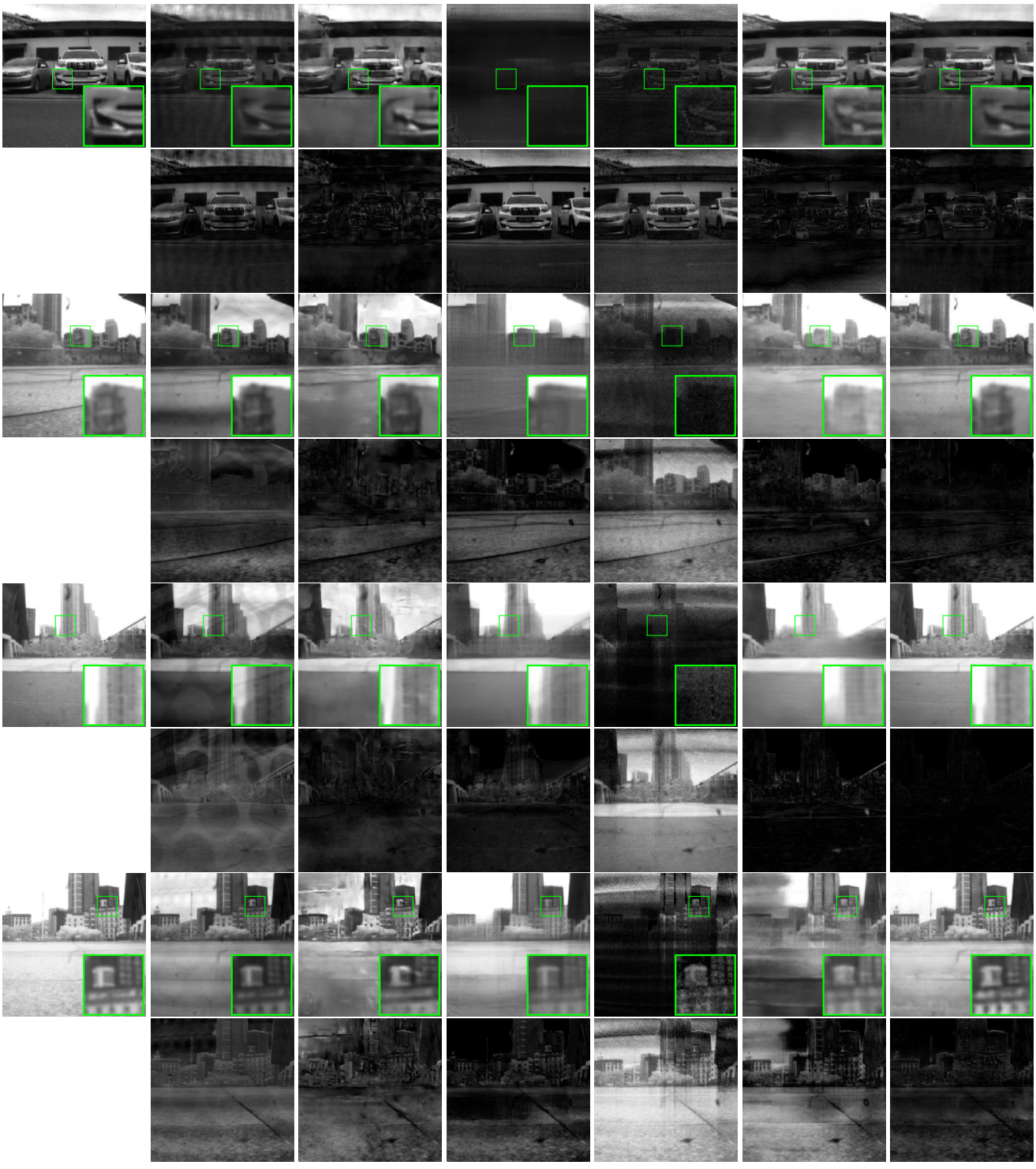

Figure 6. Qualitative comparisons of reconstructed images with corresponding error maps on the outdoor scenes under dense occlusions (1st to 2nd rows) and sparse occlusions (3rd to 8th rows).

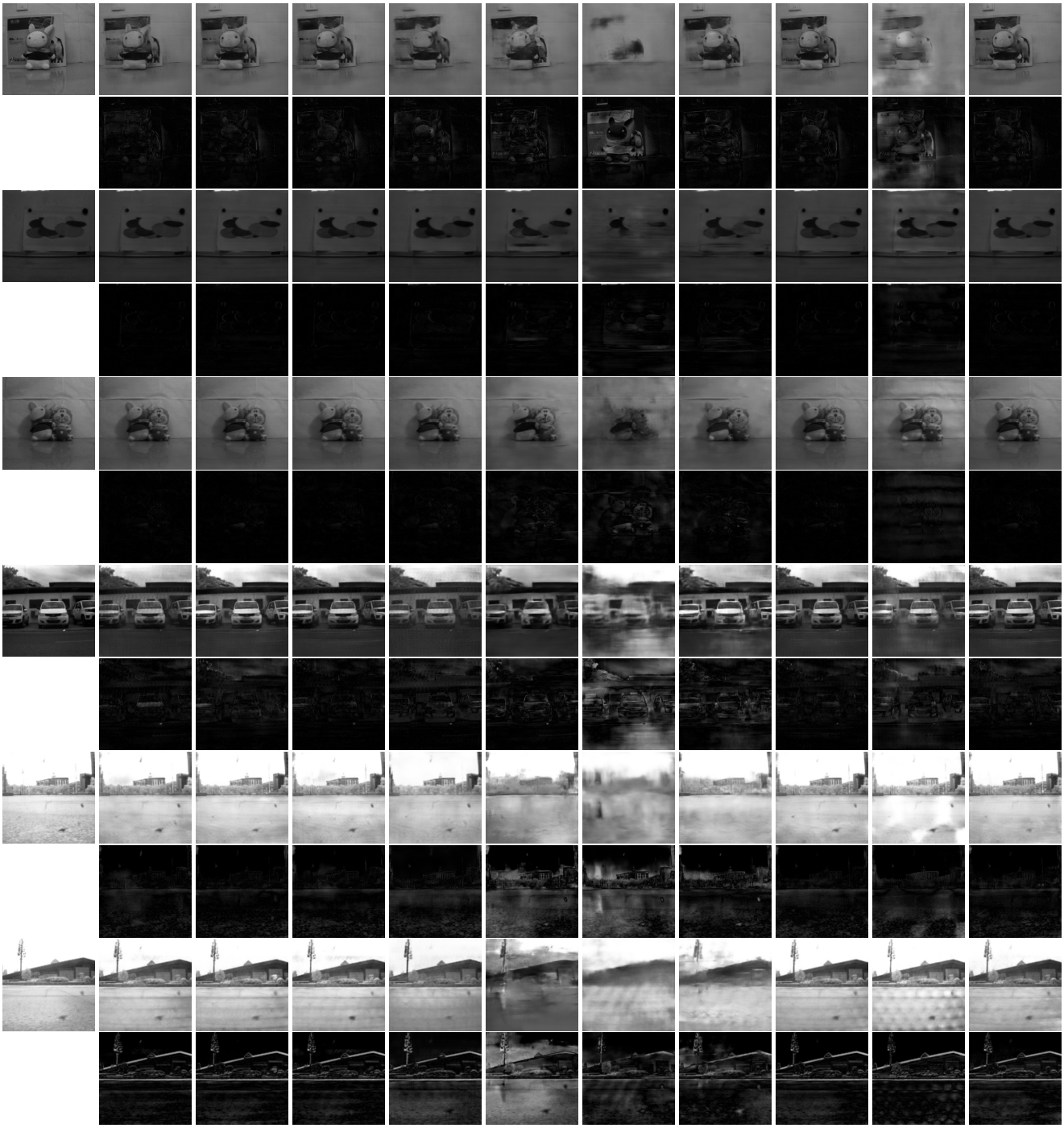

Figure 7. Qualitative ablation study of reconstructed images with corresponding error maps.