# Low Light Video Enhancement by Learning on Static Videos with Cross-Frame Attention Supplementary Material

Shivam Chhirolya
shivamchhirolya@gmail.com
Sameer Malik
sameer@iisc.ac.in
Rajiv Soundararajan
rajivs@iisc.ac.in

Department of Electrical
Communication Engineering,
Indian Institute of Science,
Bangalore, India

## 1 Architecture of Fusion Module

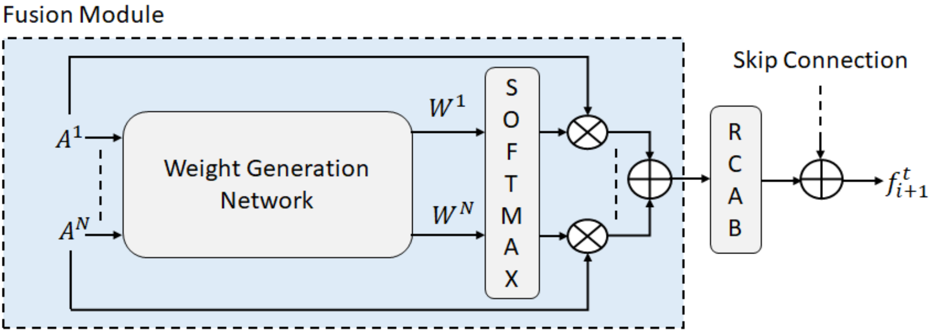

Figure 1: Block diagram illustrating our fusion module

In this section, we describe the adaptive fusion for $N$ attention maps (see Figure 1), $\{A^1, A^2, \ldots, A^N\}$. We pass each of these maps through a convolutional layer and get $W^n$ for $n = 1, 2, \ldots, N$, where each $W^n$ is a single channel map. The same map will be used to determine the weights for all the channels in the attention map. Finally, to obtain the mixing weights for $A^n$ at each location, we apply a softmax operation at each spatial location across all the $N$ maps. The resulting weights are used to fuse the attention map followed by the residual channel attention block (RCAB).

## 2 Ablation on DRV-SM Dataset

We evaluated the importance of various components of our model, in particular, cross-attention and dilated cross-attention with respect to our base line model on the DAVIS dataset in the main paper. Here we present similar results on the DRV-SM dataset using the ST-RRED quality measure in Figure 2. We see consistent improvements with respect to our ideas even in the DRV-SM dataset.

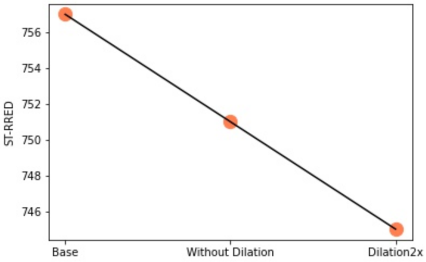

Figure 2: Ablation Evaluation on our model with DRV-SM dataset

## 3 Comparison with Data Generation based Methods

We now describe our experiments in comparing with a data generation method for training low light video enhancement [1]. SIDGAN adopts a two-step procedure to generate realistic low light videos from high quality videos in the Vimeo dataset [2]. In particular, a pair of cycle-GANs is used where the first cycle-GAN generates sensor specific long exposure video frames from which a second cycle-GAN generates short exposure video frames. Thus, paired labelled data is created which can be used to train deep network. However, we observe that each cycle-GAN generates very poor quality images in our experiments as shown in Figure 3. Thus we did not proceed with training the enhancement model. We believe that poor quality of video frames could perhaps be attributed to the larger image resolutions that we operate with when compared to SIDGAN [1]. Due to memory constraints, we are not able to experiment with larger patch sizes in their model which could potentially improve performance.

## References

[1] Danai Triantafyllidou, Sean Moran, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Low light video enhancement using synthetic data produced with an intermediate domain mapping, 2020. URL `https://arxiv.org/abs/2007.09187`.

[2] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127 (8):1106–1125, 2019.

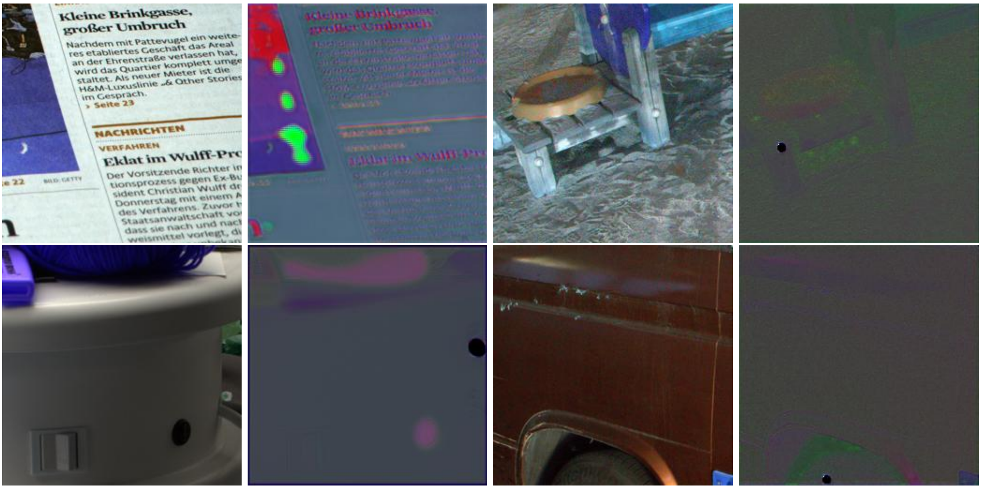

Figure 3: Examples of images generated using the SIDGAN []. First column are the ground truth images from the Vimeo dataset. Second column contains outputs from first stage of SIDGAN of transforming Vimeo ground truth to sensor specific ground truths from the DRV dataset. Third column contains ground truth images from the DRV dataset and the last column contains outputs of the second stage of the SIDGAN of transforming images from third column to low light from the DRV dataset. Note that the generated images appear unnatural.