In Figure 3 of the original paper, the merge and magnification module is composed of 2 convolutional layers and 2 sub-pixel magnification (SPM) layers [6]. Taking a feature map of $H \times W \times 64$ as input, for instance, it becomes $H \times W \times 48$ after the first convolutional layer. Then, a SPM layer reshapes it as $2H \times 2W \times 12$. Then, this feature map is further passed through the second convolutional layer and SPM layer, whose shape becomes $4H \times 4W \times 3$ ("3" denotes RGB color channels) in the end. This way, low-resolution (LR) feature map is magnified by 4 times to become a high-resolution (HR) one.

As has been discussed in Section 4.2 of the original paper, for models involving motion estimation and motion compensation (ME&MC) [1, 7], the total loss function can be described as follows:

$$\mathcal{L} = \mathcal{L}_{SR} + \lambda \mathcal{L}_{ME}, \tag{1}$$

where $\mathcal{L}_{SR}$ is adopted for super-resolution (SR) network and has been described in Section 4.1 of the original paper, $\mathcal{L}_{ME}$ is for the motion estimation subnetwork, and λ is empirically set as 0.01.

We first explain the function of the motion estimation module:

$$F_{i\to j} = (u_{i\to j}, v_{i\to j}) = ME(I_i, I_j; \theta_{ME}), \tag{2}$$

where $F_{i\to j} = (u_{i\to j}, v_{i\to j})$ denotes the optical flow field generated from input frame I_i to I_j, $ME(\cdot)$ represents the operator for calculating optical flow, and θ_{ME} stands for the corresponding parameter. The calculated optical flow $F_{i\to j}$ is later adopted for motion compensation.

We now give the specific formulation of $\mathcal{L}_{ME}$ [7]:

$$\mathcal{L}_{ME} = \sum_{i=-T}^{T} \left\| I_i^L - \tilde{I}_{0\to i}^L \right\|_1 + \alpha \left\| \bigtriangledown F_{i\to 0} \right\|_1. \tag{3}$$

where I_i^L is i_{th} LR frame, $\tilde{I}_{0\to i}^L$ represents the backward warped I_0^L according to optical flow $F_{i\to 0}$, $\bigtriangledown F_{i\to 0}$ denotes the total variation on (u, v) of $F_{i\to 0}$ as described in Equation (2), and α is also empirically set as 0.01.

Besides, as illustrated in Figure 1, we give more visual results of our model PFNL and other state-of-the-art methods like VESPCN [1], RVSR-LTD [4], MCResNet [3], DRVSR [7], FRVSR [5] and DUF_52L [2]. Except for LR frames downsampled from HR videos (shown in Figure 1(a) and Figure 1(b)), we also conduct experiments on real world videos without corresponding HR videos (taken by ourselves and shown in Figure 1(c)). We also provide source files of these images and corresponding videos in the zipped file.

References

[1] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2848–2857, July 2017.

[2] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3224–3232, June 2018.

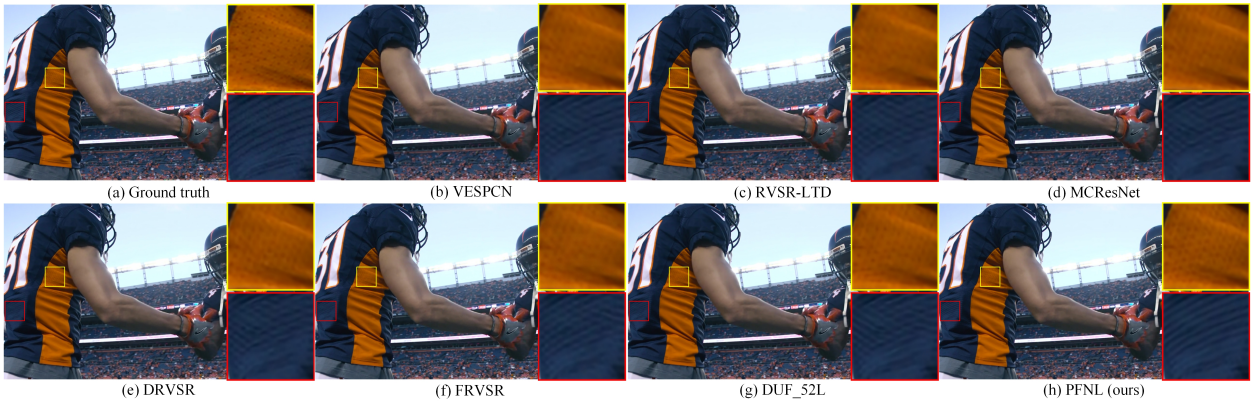
[3] Dingyi Li and Zengfu Wang. Video superresolution via motion compensation and deep residual learning. *IEEE Transactions on Computational Imaging*, 3(4):749–762, 2017.

[4] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *IEEE International Conference on Computer Vision*, pages 2526–2534, 2017.

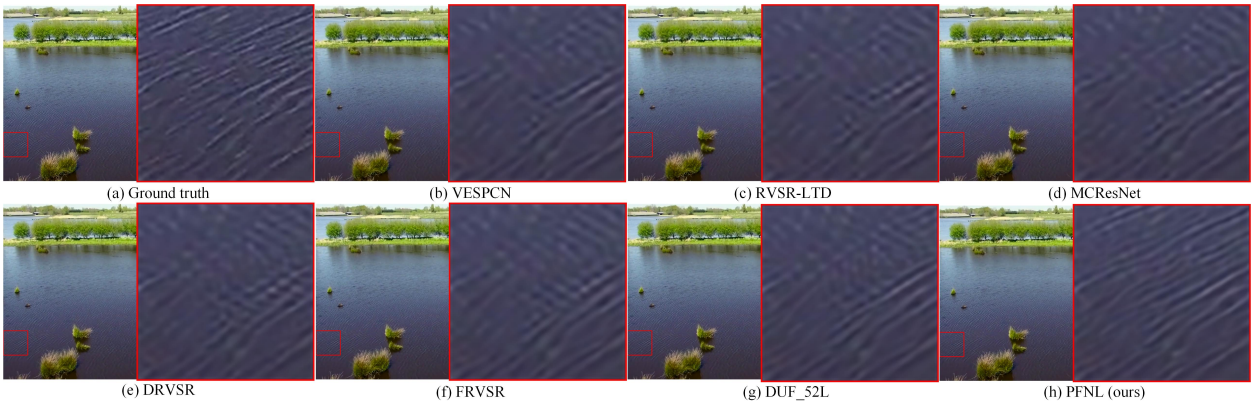
[5] Mehdi S. M Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6626–6634, June 2018.

[6] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

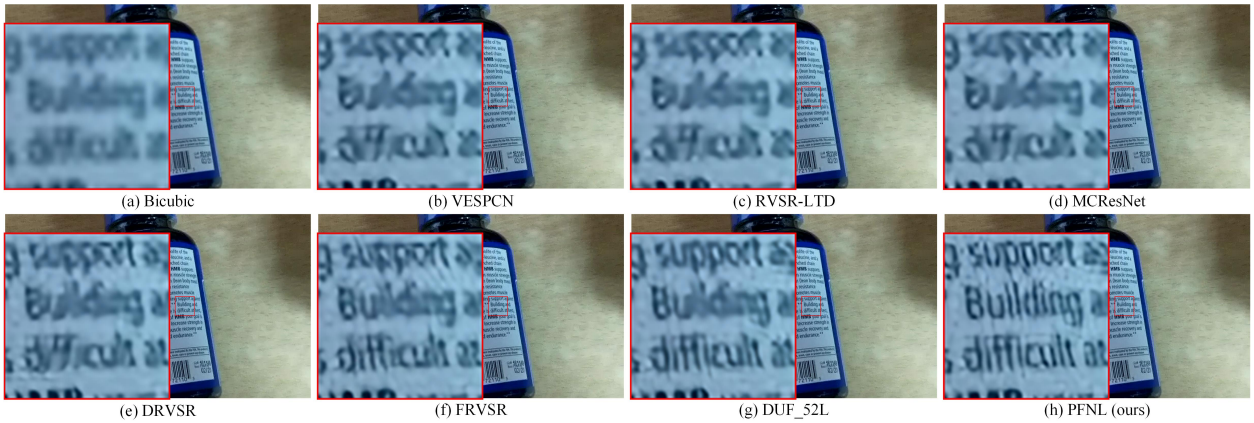
[7] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *IEEE International Conference on Computer Vision*, pages 4482–4490, 2017.



(a) This frame is from *clap*



(b) This frame is from *lake*



(c) This frame is from a real LR video rather than downsampled from an HR video.

Figure 1: Visual results of different video SR methods, for $4\times$ upscaling.