2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) | 979-8-3503-0067-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/APSIPAASC58517.2023.10317115

Generative Adversarial Network-Based Frame Interpolation with Multi-Perspective Discrimination

Quang Nhat Tran and Shih-Hsuan Yang*
Family Technology Company Limited, Vietnam
E-mail: tnquang1416@gmail.com Tel: +87-2436616333
* National Taipei University of Technology, Taiwan
E-mail: shyang@mail.ntut.edu.tw Tel: +886-2-27712171 ext. 4211

***Abstract*— Video frame interpolation plays an important role in video applications. This paper presents a new deep learning-based frame interpolation model under the generative adversarial networks (GANs) framework. The devised generator cascades three multi-scale sub-modules to capture coarse-to-fine visual characteristics. Furthermore, the generator is trained with an adequate loss function consisting of two adversarial and four reconstruction losses. Additional to the conventional discriminator that derives the spatial discrepancy, a new discriminator is employed to strengthen the temporal consistency among adjacent frames. The proposed method not only inherits the merit of GANs in processing speed but also produces high-quality video frames. Experimental results show that the proposed method achieves the most favorable image quality in PSNR and SSIM with a considerably short runtime compared to other state-of-the-art frame interpolation approaches.**

I. INTRODUCTION

Video frame interpolation is a technique that synthesizes intermediate frames from existing frames. Video frame interpolation finds many applications, including video creation, frame rate-up conversion, and video coding. Since the boom of deep learning, deep learning-based frame interpolation has caught much attention. Most frame interpolation networks are pixel-wise, and the most well-known approaches include kernel-based or optical flow-based ones. The kernel-based methods [1-3] derive texture effectively by aggregating neighbor pixels to estimate spatial kernels. Due to kernel size limitations, these approaches usually fail on large motions. The optical flow-based methods [4-7] derive texture effectively by aggregating neighbor pixels to estimate spatial kernels. Due to kernel size limitations, these approaches usually fail on large motions.

The Generative adversarial network (GAN) [8] is a deep learning data generation framework that iteratively trains adversarial networks, which are generators and discriminators. The generator synthesizes new data, deluding the discriminator, while the discriminator aims to recognize the corrected data. Based on that tactic, GAN has been applied in many images and video applications, including inpainting, random image generation, text-to-image translation, and video frame interpolation. Further improvement that conditions the data generation [9] with auxiliary data was proposed. Instead of only using latent vectors [10], labels, maps, and optical flow can be used as input. In video frame interpolation, GAN research also considers using attention maps, occlusion maps, and linear data as auxiliary inputs.

Considering quality and complexity, we introduce a GAN-based framework for frame interpolation comprising a generator and two discriminators (Fig. 1) without additional input requirements. The generator applies a three-scale architecture to capture multi-perspective features for deriving a frame from two adjacent ones. The discriminators consider

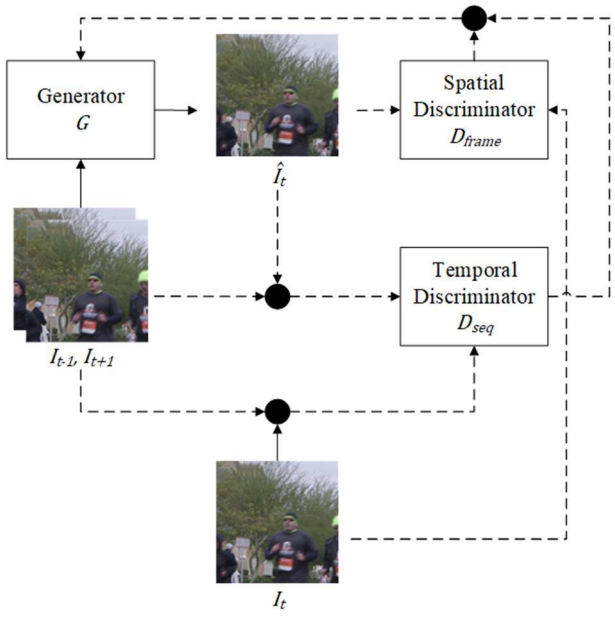

Fig. 1 The overall architecture of the proposed framework, in which two discriminators, D_{frame} and D_{seq}, facilitate a multi-scale generator G spatially and temporally. The solid and dashed arrows depict the flows of generation and discrimination, respectively.

This work was financially supported in part by the Ministry of Science and Technology, Taiwan, R.O.C. , under Grants MOST 110-2221-E-027-083-MY2.

spatial and temporal perspectives corresponding to frame-based and sequence-based models.

The main contributions of this study are summarized below:

- We design and implement a GAN-based frame interpolation framework, considering spatial correctness and temporal consistency.
- We introduce a training loss function for the generator consisting of adversarial and reconstruction loss ones to measure the reality and correctness of the output.
- We indicate each incorporated element's contribution with ablation studies and show the performance with a comparison with state-of-the-art solutions.

II. Related Works

Most recent deep learning frame interpolation solutions fall into the kernel- and optical flow-based categories. Niklaus *et al*. [1] introduced a kernel-based neural network for synthesizing intermediate frames via adaptive separable kernel estimation. Niklaus and Liu [13] manipulated optical flows and interpolation kernels with their proposed SoftSplat with forward warping operations. Besides, Liu *et al*. [4] used optical flows to copy pixels from adjacent frames for interpolation. Jiang *et al*. [5] introduced an end-to-end multi-frame interpolation network utilizing convolutional neural networks (CNNs) for jointly estimating bi-directional optical flow and occlusion mask. They also addressed the motion artifacts with visibility maps predicted by a U-Net [11]-based network.

Recently, several methods for frame interpolation have incorporated supplementary information, such as optical flows, depth maps, and occlusion maps, for high frame generation performance. Park *et al.* [7] proposed a frame interpolation network jointly estimating bilateral motion with bilateral cost volume to facilitate the optical flow estimation and dynamic filter generation network. Bao *et al.* [12] proposed an adaptive warping layer to enhance frame interpolation by integrating optical flows and interpolation kernels for motion estimation and motion compensation-driven design. From a similar point of view, Bao *et al.* [6] further used depth maps to reduce occluded objects' affection with their depth-aware projection layer. Kong *et al.* [14] introduced a pyramid frame interpolation network that refines the bilateral intermediate flow fields and features iteratively. The authors further trained IFRNet with a task-oriented optical flow distillation loss for higher performance and a geometry consistency regularization loss for enhancing structure layout. Huang *et al.* [15] introduced a frame interpolation network that utilizes intermediate optical flows with a coarse-to-fine strategy and leverages knowledge distillation from the teacher flow network.

GANs with an adversarial training strategy enable a different direction for high-quality and robust frame interpolation. GAN-based frame interpolation frameworks condition the generation to derive intermedia data by taking adjacent frames and other auxiliary data as input. Amersfoort *et al.* [16] considered a pyramid architecture to design a GAN-based frame interpolation framework. Wen *et al.* [17] proposed a multi-frame interpolation framework by concatenating two pairs of generator and discriminator aided with linear input data. Xiao and Bi [18] introduced a GANs frame interpolation framework utilizing multi-scale architecture and residual connections. Their design utilizes four pairs of generator-discriminator and incorporates attention map-based frame synthesis to enable the model to focus more on relevant regions of features based on its own estimated attention map. Tran and Yang [19] proposed a framework incorporating a two-scale generator and a discriminator to synthesize an intermediate frame from the adjacent ones.

Inspired by the previous works, the proposed method utilizes a generator for multi-perspective features and two discriminators for distinguishing spatially and temporally.

III. Proposed Methods

In this research, the proposed method aims to synthesize an intermediate frame from two given ones, as below equation:

$$\hat{I}_t = G(I_{t-1}, I_{t+1}) \tag{1}$$

Here, the target data belong to the $C \times H \times W$ vector space, in which C, H, and W represent the picture's channel, height, and width.

In the rest of this section, we describe the proposed GAN-based frame interpolation framework (Fig. 1) and explain the design for balancing generation and discrimination. We also provide a detailed description of the training loss function comprising adversarial and reconstruction loss terms.

A. Frame Interpolation Framework

Following conditional GAN-based design, the proposed method utilizes a multi-scale generator for frame interpolation and two discriminators for adversarial training. Here, the generator takes the feedback from both the discriminators, in which one considers the spatial correctness and another also pays attention to temporal consistency.

Multi-scale Generator. To explicit the multi-scale features from input and synthesize output from coarse to fine, the generator (Fig. 2) incorporates two different sub-modules deriving frames at different resolutions, as below equations:

$$\hat{I}_t^2 = G_2(I_{t-1}^2, I_{t+1}^2), \tag{2}$$

$$\hat{I}_t^1 = G_1\left(I_{t-1}^1, S(\hat{I}_t^2), I_{t+1}^1\right), \tag{3}$$

$$\hat{I}_t^0 = G_0\left(I_{t-1}^0, S(\hat{I}_t^1), I_{t+1}^0\right), \tag{4}$$

where I_t^i and $\hat{I}_t^i$ (i = 0, 1, 2) represent the existing frame and generated frame at time step t at the i^{th} level, respectively, and

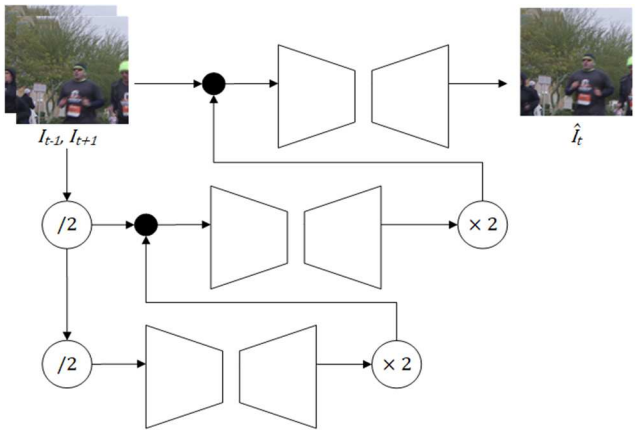

Fig. 2 The general architecture of the proposed generator.

S depicts the up-scale operation. There, the coarsest one captures the overall structure layout, and the finer one completes with more details. The finest frame $\hat{I}_t^0$ benefits from the previous inter-output $\hat{I}_t^1$ and $\hat{I}_t^2$.

All generator sub-modules share a similar backbone design following encoder-decoder architecture. Specifically, the encoder extracts features from input with a sequence of convolutional layers, batch normalizations, and Parametric Rectified Linear Unit (PreLU) [20] activation functions. Here, we use PReLU at the encoder side to reduce feature loss in the extraction process with convolution operations. Thus, the models become more robust in the deep learning environment. Furthermore, the generator is incorporated with skip connections for training efficiency enhancement and recovering spatial information by explicitly copying features between layers. Accordingly, the decoder synthesizes the output with transposed convolutional layers, batch normalizations, Dropout layers, and Rectified Linear Unit (ReLU) activation functions.

Spatial- and Temporal-Discriminators. A generated frame should contain corrected and consistent content with the adjacent ones. Therefore, we utilize two discriminators for classifying the output spatially and temporally. The first network evaluates frames individually (5), while the second considers a patch of three consecutive frames for temporal consistency (6). In this way, discrimination produces better feedback for the generation and facilitates the generation. Besides, this design improves the discrimination efficiency that helps balance the capacity and performance between generated and discriminated operations for more efficient training.

$$d_{spatial} = D_{spatial}(\hat{I}_t^0) \quad (5)$$

$$d_{temporal} = D_{temporal}(I_{t-1}^0, \hat{I}_t^0, I_{t+1}^0) \quad (6)$$

The two spatial and temporal discriminators share a similar backbone architecture (Fig. 3) that sequences convolutional layers, Batch normalizations, PReLU activation functions, and a sigmoid activation function.

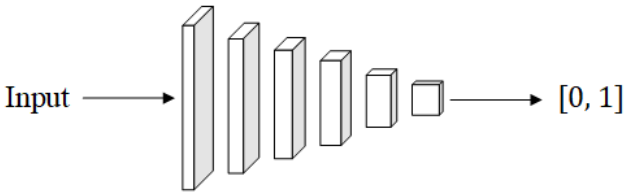

Fig. 3 The general architecture of the proposed discriminators.

B. Training Loss Function

Following the GAN-based design [8], we apply the minimax approach for training the proposed framework. As the below equation, the generator tries to minimize the classification accuracy, the term $(1 - D(x))$, with the generated data. Meanwhile, the discriminator attempts to maximize it.

$$\min_G \max_D E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_g(x)}[\log(1 - D(x))] \quad (7)$$

We train the two discriminators by measuring classification accuracy with the Binary Cross Entropy (BCE) loss function. This function is also applied for measuring the video frame degradation as the adversarial loss term besides the reconstruction one, as (8), where λ_{adv} and λ_{rec} depict the non-negative weights. In practice, we apply 0.00005 for λ_{adv} and 1.0 for λ_{rec}.

$$\mathcal{L} = \lambda_{adv} \times \mathcal{L}_{adv} + \lambda_{rec} \times \mathcal{L}_{rec} \quad (8)$$

The adversarial loss term $\mathcal{L}_{adv}$ measures the feedback from the spatial and temporal discriminators, as (9), and allows the generator considers them for learning.

$$\mathcal{L}_{adv} = \mathcal{L}_{spatial} + \mathcal{L}_{temporal} \quad (9)$$

For reconstruction loss function, we refer to FI-DUSGAN [19], incorporating $\mathcal{L}_2$, multi-scale structural similarity (MS-SSIM) loss [21] $\mathcal{L}_{MS\text{-}SSIM}$, and gradient difference loss [22] $\mathcal{L}_{GDL}$ for pixel-wised generation correctness, low motion blur, and high friendliness in human vision. In this work, we add $\mathcal{L}_{census}$ to enhance the generation learning for balancing the performance between the generator and discriminators. This function measures the soft Hamming distance between ternary census transformed images with a patch size of 7×7 [23] to facilitate the generation with a more reliable constancy assumption based on the additive and multiplicative illumination changes compensation of the census transform. The proposed function is described below.

$$\mathcal{L}_{rec} = \mathcal{L}_2 + \mathcal{L}_{census} + \mathcal{L}_{MS-SSIM} + \mathcal{L}_{GDL} \quad (10)$$

IV. EXPERIMENTS

In this section, we first describe the experimental setup, including the dataset and parameters for training and implementation. Then, we evaluate the generation performance by an ablation study and a comparison with the state-of-the-art in terms of peak signal-to-noise ratio (PSNR) for pixel-by-pixel comparison and SSIM for perception-based similarity.

A. Experimental Setup

In this experiment, we train the proposed framework with the temporal frame interpolation subset of the Vimeo-90k dataset [24]. The dataset consists of several 480×256-frame patches containing three consecutive frames. Besides, we further pre-process the training dataset with random cropping and flipping operations. The target training dataset includes 51,313 128×128-frame-patches, and the testing dataset has 3,782 480×256-frame-ones.

We use PyTorch to develop the models and manipulate the training with the Adam optimizers with the recommendation parameters for implementation. Besides, we set the training batch size as 32 and the initial learning rate as 0.001 while decaying it after every 20 training epochs until reaching 10^{-7}.

We perform the experiment on the NVIDIA Tesla V100 GPU environment.

B. Ablation Study

In this ablation study, we compare various versions of the proposed framework to demonstrate the contributions of each component. As in Table I, the proposed design (5) produces the highest quality with 37.20dB in PSNR and 0.940 in SSIM. This version confirms the contribution of the modifications and the efficiency in balancing between generation and discrimination.

Table I. Ablation study on architecture and training loss function, in which the best version is highlighted in bold.

Version	Description	Loss	PSNR	SSIM
(1)	3-scale generator only	$\mathcal{L}_2$-based	36.85	0.940
(2)	2-scale-generator and seq-discriminator	$\mathcal{L}_2$-based	36.73	0.941
(3)	3-scale-generator and seq-discriminator	$\mathcal{L}_1$-based	35.75	0.928
(4)	3-scale-generator and seq-discriminator	$\mathcal{L}_2$-based	36.39	0.939
(5) Proposed	**3-scale-generator and 2 discriminators**	**$\mathcal{L}_2$-based**	**37.20**	**0.940**

Table II. Ablation study on weights of the training loss function, in which the proposed version is highlighted in bold.

Version	λ_{adv}	λ_{rec}	PSNR	SSIM
(1)	0.0001	0.5	36.84	0.938
(2)	**0.00005**	**1.0**	**37.20**	**0.940**
(3)	0.0	1.0	36.85	0.940

According to Table I, a good design GAN-based framework having efficient adversarial training will allow us to simplify the network, such as the two-scale and the three-scale one of (2) and (1), respectively. Versions (3) and (4) produce lower performance due to the imbalance between generation and discrimination. Here, the generator is more complex, and the discriminator does not have enough capability to keep the training effective. Therefore, the incorporated frame-based discriminator helps balance the generation and discrimination, enabling the proposed framework to achieve higher performance. From a different point of view, we consider replacing $\mathcal{L}_2$ with $\mathcal{L}_1$ for training the generator. However, the low-quality output of (3) shows that the mean square error-based loss function is more compatible with the proposed architecture and training loss function. In our observation, incorporating one more discriminator into (2) causes the vanishing gradients problem, and the generator fails to learn. The training problem confirms the suitability of our model designs with the applied training loss function. Besides, our study on weights of the training loss function (Table II) shows the effectiveness of the selected ones for the proposed function.

C. Comparison to the state-of-the-arts

In this section, we perform a comparison between DVF [4], SepConv [1], FI-MSAGAN [18], BMBC [7], FI-DUSGAN [19], and the proposed framework. For comparison, we retrain the compared models using a re-implemented version of FI-MSAGAN [25] and official implementations of others. Here, the proposed generator achieves the best PSNR and second-best SSIM values (Table III) with the Vimeo-90k dataset within a short runtime (9.1 ms). Overall, GAN-based methods can

Table III. Comparison between the frame interpolation methods, in which the best and second-best values are highlighted in bold and underlined, respectively.

Methods	PSNR (dB)	SSIM	Runtime (ms)
DVF	31.92	0.658	51
SepConv	35.97	0.913	115
FI-MSAGAN	35.99	0.893	26
BMBC	36.74	**0.948**	686
FI-DUSGAN	<u>37.02</u>	<u>0.940</u>	9
The proposed	**37.20**	<u>0.940</u>	9.1

(a) PSRN: 37.93dB; SSIM: 0.966

(b) PSRN: 38.40dB; SSIM: 0.952

Fig. 4 The generated output of the proposed framework.

Fig. 5 The generated output of the proposed framework (PSRN: 34.78dB; SSIM: 0.962).

achieve high-quality output with lower complexity owing to the efficient adversarial training. In our comparison, the proposed framework outperforms FI-DUSGAN by 0.18dB in PSNR and FI-MSAGAN by 1.21dB in PSNR and 0.47 in SSIM.

As shown in Fig. 4, the proposed framework synthesizes interpolated frames well in various scenarios, including frames with high contrast details with low to medium changes. On the other hand, it still fails to handle some challenging scenarios, particularly those with fast and sudden motions between frames, such as the moving cars in Fig. 5.

V. CONCLUSION

We introduced a frame interpolation framework utilizing three adversarial networks, including a three-scale generator and two discriminators. The devised generator captures coarse-to-fine visual characteristics by three-scale sub-modules with an adequate combination of loss functions. Besides a discriminator considering the quality of generated data, a temporal one is employed to guarantee consistency among adjacent frames. Thus, the proposed framework effectively balances the adversarial relationship to ensure optimal performance. Our ablation study confirmed the contribution of each designed element. The proposed design improves the PSNR by 0.35 dB without increasing complexity compared to the generator-only version. The proposed framework can derive high-quality frames in various scenarios, and it outperforms other GAN-based solutions, such as FI-MSAGAN (+3.4%) and FI-DUSGAN (+0.5%), with a similar complexity requirement. In future work, we will consider tackling more challenging scenarios, such as videos with sudden motions.

REFERENCES

[1] S. Niklaus, L. Mai and F. Liu, "Video Frame Interpolation via Adaptive Separable Convolution," in IEEE International Conference on Computer Vision (ICCV), Venice, 2017.

[2] T. Ding, L. Liang, Z. Zhu and I. Zharkov, "CDFI: Compression-Driven Network Design for Frame Interpolation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, 2021.

[3] X. Cheng and Z. Chen, "Multiple Video Frame Interpolation via Enhanced Deformable Separable Convolution," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 7029-7045, 2022.

[4] Z. Liu, R.A. Yeh, X. Tang, Y. Liu and A. Agarwala, "Video Frame Synthesis Using Deep Voxel Flow," in International Conference on Computer Vision (ICCV), Venice, 2017.

[5] H. Jiang, D. Sun, V. Jampani, M. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018.

[6] W. Bao, W-S. Lai, C. Ma, X. Zhang, Z. Gao, M-H. Yang, "Depth-Aware Video Frame Interpolation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[7] J. Park, K. Ko, C. Lee and C-S. Kim, "BMBC: Bilateral Motion Estimation with Bilateral Cost Volume for Video Interpolation," in European Conference on Computer Vision, 2020.

[8] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in Neural Information Processing Systems (NIPS), 2014.

[9] M. Mirza and S. Osindero, Conditional Generative Adversarial Nets, arXiv preprint arXiv:1411.1784v1, 2014.

[10] A. Radford, L. Metz and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434v2, 2016.

[11] O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, arXiv preprint arXiv:1505.04597, 2015.

[12] W. Bao, W-S. Lai, X. Zhang, Z. Gao and M-H. Yang, "MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 3, pp. 933-948, 2021.

[13] S. Niklaus and F. Liu, "Softmax Splatting for Video Frame Interpolation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[14] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, Ch. Wang and J. Yang, IFRNet: Intermediate Feature Refine Network for Efficient Frame Interpolation, arXiv prePrint arXiv:2205.14620, 2022.

[15] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-Time Intermediate Flow Estimation for Video Frame Interpolation," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, 2022.

[16] J.V. Amersfoort, W. Shi, A. Acosta, F. Massa, J. Totz, Z. Wang and J. Caballero, Frame Interpolation with Multi-Scale Deep Loss Functions and Generative Adversarial Networks, arXiv preprint arXiv:1711.06045, 2019.

[17] S. Wen, W. Liu, Y. Yang, T. Huang and Z. Zeng, "Generating Realistic Videos From Keyframes With Concatenated GANs," IEEE Transactions on Circuits and Systems for Video Technology, vol. 29, no. 8, pp. 2337 - 2348, 2019.

[18] J. Xiao and X. Bi, "Multi-Scale Attention Generative Adversarial Networks for Video Frame Interpolation," IEEE Access, vol. 8, pp. 94842-94851, 2020.

[19] Q.N. Tran and S-H. Yang, "Video Frame Interpolation via Down-Up Scale Generative Adversarial Networks," Computer Vision and Image Understanding, vol. 220, 2022.

[20] K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in IEEE International Conference on Computer Vision (ICCV), Santiago, 2015.

[21] H. Zhao, O. Gallo, I. Frosio and J. Kautz, "Loss Functions for Image Restoration With Neural Networks," IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 47-57, 2017.

[22] M. Mathieu, C. Couprie and L. LeCun, "Deep multi-scale video prediction beyond mean square error," in International Conference on Learning Representations (ICLR), 2016.

[23] S. Meister, J. Hur and S. Roth, "UnFlow: Unsupervised Learning of Optical Flow With a Bidirectional Census Loss," in AAAI Conference on Artificial Intelligence, 2018.

[24] T. Xue, B. Chen, J. Wu, D. Wei and W.T. Freeman, "Video Enhancement with Task-Oriented Flow," International Journal of Computer Vision, vol. 127, no. 8, pp. 1106-1125, 2019.

[25] Q. N. Tran, "A Reimplementation of FI-MSGAN Using PyTorch," 2021. [Online]. Available: https://github.com/tnquang1416/FI-MSGA.