Multi-Scale Attention Generative Adversarial Networks for Video Frame Interpolation

JIAN XIAO\textsuperscript{1} AND XIAOJUN BI\textsuperscript{2}

\textsuperscript{1}College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
\textsuperscript{2}School of Information Engineering, Minzu University of China, Beijing 100081, China

Corresponding author: Xiaojun Bi (bixiaojun@hrbeu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51779050.

Abstract Video frame interpolation is a fundamental task in computer vision. Recent methods usually apply convolutional neural networks to generate intermediate frames with two consecutive frames as inputs. But sometimes existing methods fail to handle with complex motion and long-range dependencies. In this paper, a multi-scale dense attention generative adversarial network is proposed. First, a multi-scale generative adversarial framework is established for video frame interpolation. Generators from coarse to fine can better combine global and local information. Second, an attention module introduced to generator makes network accurately focus on moving objects. Third, a sequence discriminator is designed to improve the ability of capturing spatial and temporal consistency in frame sequence. Experimental results of ablation study prove the effectiveness of our three contributions. And results on several datasets demonstrate that our approach attains higher performance and produce more photo-realistic in-between frame comparing with previous works.

Index Terms Video frame interpolation, generative adversarial networks, multi-scale pyramid, spatial and temporal consistency, sequence discriminator.

I. INTRODUCTION

Video frame interpolation is a fundamental issue in computer vision and aims to synthesize plausible intermediate frames between any two adjacent frames. This is a challenging problem requiring a generative framework that can model motion information, and generate both spatially and temporally consistent frames. Video frame interpolation can be used in numerous applications, such as video frame rate conversion [1], slow-motion video generation [2] and video compression [3]. It helps surmount the temporal constraint of camera sensors and has drawn increasing attention in image and video processing community.

Video frame interpolation is a complicated problem due to three reasons: (1) The variability of real world scene makes the video sequences modeling considerably difficult. (2) Large motion generally appears in videos. It is more troublesome to capture temporal and spatial consistency in that case. (3) The video interpolation framework could not generate realistic frames with accurate content when local occlusion occurs in the video.

Most existing video interpolation methods usually use convolutional neural networks (CNN). These methods can be divided into two categories: based on interpolation convolution kernels estimation [4], [5] and based on optical flow estimation [2], [6], [7]. The first technique combines motion estimation and pixel synthesis into a single process, which produces a convolution kernel for each pixel. It can capture both the motion information between the original consecutive frames and the coefficients for pixel synthesis via a convolutional neural network. The other technique estimates optical flow between input frames firstly, and then interpolate intermediate frames depend on the estimated dense correspondences. The accuracy of optical flow estimation directly affects the quality of the generated frame. However, these two kinds of CNN-based techniques are essentially indirect methods to synthesize intermediate frames. On the one hand, they will produce unrealistic and blurry results if the intermediate process fail. The estimated optical flow or kernel is commonly not correct when the motion regions under the case of occlusion, blur and sudden illumination change. And on the other hand, the computation of optical flow or kernel for each pixel requires considerable time consumption and memory overhead, which significantly escalates the difficulty.
of video frame interpolation. Recently, generative adversarial networks (GANs) have achieved great success in the field of image and video generation [8], [12], [48]. GANs contain a generator network and a discriminator network, which are trained against each other. Through adversarial learning, generator can achieve satisfactory results.

Building an effective frame interpolation model of natural images is hard due to video sequence are complex and high dimensional. It is easy to get blurred results and artifacts when processing high-resolution image, because there are different sizes of objects in the image. Although we can contact multiple convolution operations to expand the receptive field, the problem of long-distance dependence will become more prominent [15], [49]. The proposed method utilizes multi-scale structure that can more accurately model global information and local features with different size level images as input. The effective image patch size become smaller as going up the multi-scale pyramid, and the network can process more elaborate image local feature.

In this paper, a novel and explicit video frame interpolation method is proposed which can relieve the calamity caused by mistaken intermediate process. Inspired by recent advances in generative adversarial networks [8]–[11] for image synthesis [12]–[14] and video prediction [15]–[17], we introduce a frame interpolation framework using multi-scale dense attention generative adversarial networks, i.e., FI-MSAGAN. FI-MSAGAN is an effective end-to-end trainable fully convolutional neural network, manipulating two consecutive frames at arbitrary resolution and directly producing the high-quality intermediate frame. The proposed FI-MSAGAN progressively reconstruct intermediate frames in a coarse-to-fine manner. At each scale, the frame generator uses residual blocks [18], [19] and skip connection to form a dense network structure. And frame discriminators judges whether the input frame patches are fake or true. It helps the model to capture fine details and texture structure. In order to make the network focus on large moving objects and handle with complex motion, we introduce an attention module for generator. In addition, a sequence discriminator is designed to provide feedback signal for generator to capture the temporal consistency between video frames. The training process of FI-MSAGAN is end-to-end and guided by a comprehensive loss function that contains a unique multi-scale frame adversarial loss, a sequence adversarial loss, a perceptual loss and a reconstruct loss. This paper mainly consists of three contributions condensed as follows:

1. We propose an efficient multi-scale dense attention generative adversarial networks (FI-MSAGAN) for video frame interpolation. Our method can utilize global and local information to produce more realistic intermediate frame directly without optical flow estimation.

2. We design a dense attention generator for better reconstruct temporal and spatial consistency of video sequences. The generator consists of synthesis module and attention module. Attention map obtained from attention module makes generator adaptively focus on the dynamic areas.

3. We construct a sequence adversarial loss through a sequence discriminator. And frame discriminators recover high-frequency structure. Ultimately, the total loss of the network includes four terms: multi-scale frame adversarial loss, sequence adversarial loss, feature perception loss, and reconstruction loss. It is adopted to further improve the quality of the interpolated frame.

II. RELATED WORK

Many advanced approaches for video frames interpolation explicitly or implicitly presume consistent motion between video sequences. Conventional video frame interpolation approaches usually consist of two stages: motions estimation and pixel synthesis. Phase-based method, optical flow and local convolution kernel can be used to capture motion consistency. Mahajan et al. [21] holds that a given pixel in the interpolated frames traces out a path in the original video sequences. Based on the intuitive idea, they copy and move pixel gradients from the inputs to the interpolated frames along this path. Meyer et al. [1] presents an efficient method to computing in-between frames by simple per-pixel phase information propagation across the levels of multi-scale pyramid, but it fails gracefully in case of large appearance changes.

Optical flow prediction is the key technology of motions estimation and determines the accuracy of the interpolation results. However, they are often challenged by large motions and brightness change [22]–[24]. Deep learning has make enormous success and dramatic developments in image and video recognition. Most state-of-the-art optical flow models use deep learning [25], which suggest that CNNs can understand motion information between frames. In order to get better results, many researchers merge optical flow estimation and video interpolation frame in a single model [2], [6], [7], [26]–[28]. Liu et al. [6] designs a deep network with a voxel flow layer to synthesize video frames by flowing pixel values from input video volume. The DVF is trained in an unsupervised manner and can be easily extended to extrapolating video. Jiang et al. [2] first uses a flow computation U-Net to estimate the bi-directional flow, which is then linearly fused to get rough intermediate flow. And then, they use another flow interpolation U-Net to refine the approximate flow and visibility maps. Finally, synthesizing intermediate frames by applying the visibility maps to the warped images. Niklaus and Liu [27] apply pixel-wise contextual information extracted by a pre-trained network to estimated bidirectional flow, and uses a frame synthesis network to produce the interpolated frame in a context-aware fashion. Zhang et al. [7] uses a 3D U-Net feature extractor to excavate spatio-temporal context and rebuild texture, and a coarse-to-fine architecture to improve optical flows estimation. Li et al. [28] proposes a lightweight network to estimate optical flow at feature level and introduce a new sobolev loss achieve better results.
Different from optical flow based models, Niklaus et al. [4] develops a network (AdaConv) to estimate spatially-adaptive convolution kernel for each pixel. However, the memory demand increases quickly while the model needs large kernel to deal with large motion. Niklaus et al. [5] improves AdaConv by using a pairs of 1D kernels to approximate regular 2D ones. The use of local separable convolution kernels significantly reduces the parameters of the model. There are several papers [29], [30] that use CNNs to directly produce intermediate frames, yet the models are simple and hard to generate reasonable results.

After the invention of generative adversarial networks [8], many researches applied this framework to generate images in the context of image-to-image translation [12], [31], super resolution [32], and video prediction [15]–[17], [48]. Mathieu et al. [15] first applies adversarial learning to video prediction, which employ an image gradient loss in a multi-scale architecture and obviously reduce blurring artifacts. On the other hand, many works have improved performance of GANs [13], [33], [34], [36]. WGAN [33] and LSGAN [36] introduce a new discriminator loss function to relieve training instability. BigGAN [13] and ProGAN [34] allow generator to map noise to high-resolution and realistic images. Inspired by this, our network uses multi-scale generative adversarial strategy to produce middle frame in an explicit manner.

III. PROPOSED APPROACH

In this section, we expatiate our method of video frame interpolation. First, we define the problem and give some notations. Then, we introduce our proposed multi-scale frame interpolation networks in details. At last, we show several types of loss functions used in the networks.

A. PROBLEM DESCRIPTION

Our goal is to synthesize the intermediate frame between two successive frames \(I_1, I_2 \in \mathbb{R}^{H \times W \times C} \), \(H, W, C\) are the height, width and channel number of the frame, respectively. We denote the generated in-between frame as \(I_s\) with the same size of \(I_g\), which is real frame used in training phase and unavailable during test.

In this work, we take advantage of generative adversarial networks to achieve more effective video frame interpolation task. The generator receives consecutive frames \(I_1\) and \(I_2\) as inputs and find a map to synthesize the intermediate frame \(I_s\):

\[
I_s = G(I_1, I_2)
\]

(1)

The discriminator distinguishes real frames from fake frames produced by the generator and provide the update signal for \(G\).

B. Fi-MSAGAN STRUCTURE FOR VIDEO FRAME INTERPOLATION

Our proposed method adopts a multi-scale structure, which consists of several generators and discriminators with different size of images as inputs. The multi-scale structure can better combine global information and local details for photo-realistic frame generation. In order to preserve the spatial and temporal consistency of the generated video, we introduce a dense attention generator. Moreover, we use both a frame discriminator and a sequence discriminator to distinguish fake data from real. The generators and discriminators as follows.

1) MULTI-SCALE DENSE ATTENTION GENERATOR

We construct the proposed video frame interpolation method using a multi-scale pyramidal structure as shown in Figure 1. There are four scale levels in the network and each level of generator is a subnetwork with several residual blocks, denote as \(G_i, i = 1, \ldots, 4\). Our generators takes original scale or downsampling version of \(I_1\) and \(I_2\) as inputs and synthesize the interpolated frames via a gradual approach. The generated frames of each level are obtained as:

\[
I^*_i = \begin{cases}
  G_i(I_1^i, I_2^i), & i = 1 \\
  G_i(I_1^i, I_2^i, U I_{i+1}^{i-1}), & i = 2, 3, 4
\end{cases}
\]

(2)

where \(U\) denotes upsampling. At the coarse level, we feed the downsampling frames \(I_1^1\) and \(I_2^1\) of size \(H/2^{i-1} \times W/2^{i-1}\) into \(G_i\). However, at other finer level, the inputs of \(G_i\) also contains the upsampled output of from \(G_{i-1}\). In practice, we add output of \(G_{i-1}\) between the frames \(I_1^i\) and \(I_2^i\) so that generator at finer level can capture temporal coherence.

The entire generator network is a cascade of sub-networks with common structure at each scale. Generator \(G_i\) consists of a synthesis module \(S_i\) and an attention module \(A_i\). The structure of \(G_i\) is described in Figure 2. For synthesis module, as shown in the blue rectangle in Figure 2, the input layer uses a convolution kernel with \(5 \times 5\) to obtain a larger receptive field. The number of residual blocks in \(S_i\) increases with the growth of input images scale. We adopt pre-activation residual block with \(3 \times 3\) kernel size and stride 1. All convolution operations output 64 feature maps, but the final layer outputs a frame with tanh activation. We utilize skip connections between the input layer feature map/shallow layer maps and the last residual block to maintain information extracted from shallow layer in the form of a dense connection.

Recent works have achieved great success in image-to-image translation [12], [31], [37], [38]. These works adopt similar architecture as [39] for generator, which contain stride-2 convolutions for down-sample, several residual blocks for domain feature space translation and fractionally-strided convolutions with stride 1/2 for up-sample. However, inspired by [15], we remove down-sample and up-sample convolution layers and only use a few residual blocks in our multi-scale generator architecture, both synthesis module \(S_i\) and an attention module \(A_i\). For each scale, the size of feature map remains unchanged in information flow. Our generator structure can avoid two disadvantages of previous works: (1) partial information loss of shallow layers feature maps caused by down-sampling convolutions and (2) up-sampling convolutions result in checkerboard effect. These generators can make use of all the information of different scale images to produce more realistic intermediate frame. The whole
FIGURE 1. Overview of our multi-scale frame interpolation network. The left side represents the multi-scale generator in yellow block and the right side represents the multi-scale frame discriminator. \( G_i \) is composed of some residual modules. \( G_1 \) is feed downsampling version of original frames as inputs. Generators at other finer level receive the output of the previous generator and downsampling version of original frames as inputs.

The generation process is from coarse to fine so that the network can better model global information and local feature.

Video frame synthesis is a quite complicated and challenging task due to large and complex motion or occlusions. Effective video interpolation algorithm should be able to accurately focus on moving objects. However, a network with a series of convolution operations could not achieve the above property. Because traditional convolution structure only account for short-range dependencies, limited by the kernel size. Recent researches show that, motivated by human perception procedure, attention mechanism has an advantage in computer vision, e.g. image classification [40], image-to-image translation [41], [42], video classification [20]. Rather than processing a single image or a sequence using local information, attention allows the network to focus on the most relevant part of features as needed.

We produce an attention map by inserting an attention module to solve the above issues. As shown in the purple rectangle in Figure 2, Attention module \( A_i \) help the generator to capture long-range dependencies so that generator \( G_i \) can adaptively focus on the dynamic areas. Attention module is slightly different from synthesis module. We remove skip connections and the final activation function is sigmoid. The attention map is continuous value between [0;1].

For example at scale \( i \), we utilize attention map \( am_i \) and input frame \( I_i \) to create the final finer intermediate frame:

\[
I_s = am_i \otimes I_s + (1 - am_i) \otimes I_1
\]  

where \( \otimes \) denote element-wise produce. The first part produces the intermediate frame, while the second part corresponds to pixels warped from the previous frame. Our attention module can effectively recognize the key region of moving objects in the frame sequence and keep the static background perfectly.

2) MULTI-SCALE FRAME DISCRIMINATOR AND SEQUENCE DISCRIMINATOR

Our network is a multi-scale structure, so we provide each generator with a corresponding frame discriminator. In other words, the whole network incorporates many pairs of generators and discriminators that can process different size input images. These GANs models generate the final intermediate frame from coarse to fine. Frame discriminators with different scales have different receptive fields compared with original size images. The coarsest scale frame discriminator has the largest receptive field and better global view, which can guide the generator to generate globally consistent frames. While the frame discriminator at the top level is adept in guiding the generator to produce more exquisite detail information. At the same time, it makes it easier to training the generator easier and more stability.

The architecture of frame discriminator \( D_i \) is similar to patchGAN [12], shown in Figure 3. It is generally known that L1 loss can merely capture the low frequency information for
image and video generation task. In our proposed method, we apply patchGAN to distinguishing image local patch so
that the frame discriminator can model high frequencies.
In order to make frame discriminator at different level focus
on image region with diverse size, $D_i$ from coarse to fine
contains 2 to 5 convolution layers. The kernel size is $3 \times 3$
used in all convolutional layers, and stride is 2 except for the
first layer is 1. The feature maps are 64, but the last layer
averages all local responses to provide the ultimate output
of $D_i$.

Temporal consistency of generated intermediate frames is
crucial for video frame interpolation. Therefore, we design a
sequence discriminator $D_S$ to guide the generator to capture
spatial and temporal consistency in video. The structure of
sequence discriminator is similar to traditional discriminator
[43], shown in Figure 4. We only adopt the sequence
discriminator at finest scale. The consecutive two real frames
and the generated intermediate frames compose a fake
sequence, $F_{seq} = contact(I_1, I_S, I_2)$, and $D_S$ distinguishes
real sequence from fake sequence.

C. LOSS FUNCTIONS

Our multi-scale video frame interpolation approach is based
on GANs [8] which optimizes the following objective functions:

$$
\min_{G} \max_{D} \mathbb{E}_{x \sim \mathcal{P}_{data}(x)} \left[ \log D(x) \right] + \mathbb{E}_{x \sim \mathcal{P}_{model}(x)} \left[ \log (1 - D(x)) \right]
$$

(4)

The proposed method uses two kinds of discriminator which
corresponding to two adversarial losses: 1) multi-scale frame
adversarial loss; 2) sequence adversarial loss. Moreover,
we adopt two reconstruction losses: 3) frame reconstruction
loss; 4) feature perceptual loss. So our objective contains four
parts in total.

We use multi-scale frame discriminators and a sequence
discriminator, so there are two kinds of adversarial loss. The
adversarial loss for frame discriminator at level $i$ as formula (5).
We Maximum it to optimize $D_i$ and minimum it to optimize $G_i$:

$$
\mathcal{L}_{GAN}^i = \mathbb{E}_{I_i \sim \mathcal{P}_{data}} \left[ \log D_i(I_i) \right] + \mathbb{E}_{I_i \sim \mathcal{P}_{model}} \left[ \log (1 - D_i(I_i)) \right]
$$

(5)
where $I^i_r$ is the real intermediate frame and $I^i_s$ is synthesized by $G_i$ as formula (2). The multi-scale frame adversarial loss $L_F$ is the weighted sum of all scale losses:

$$L_F = \sum_{i=1}^{4} \lambda_i L^i_{GAN}$$  \hspace{1cm} (6)$$

Sequence discriminator $D_S$ output a probability that the input is a real frame sequences. The sequence contains three consecutive frames of the original size, i.e. $Fseq = contact(I_1, I_S, I_2)$ and $Rseq = contact(I_1, I_r, I_2)$. The sequence adversarial loss is only employed at top level:

$$L_S = \mathbb{E}_{p_{data}} \left[ \log D_S (Rseq) \right] + \mathbb{E}_{p_{model}} \left[ \log (1 - D_S (Fseq)) \right]$$  \hspace{1cm} (7)$$

We use a pixel-wise reconstruction loss in L1-norm, which can produce sharper results than MSE, the expression as follow:

$$L_f = \mathbb{E}_{p_{data}} \left[ \| I_s - I_r \|_1 \right]$$  \hspace{1cm} (8)$$

Inspired by perceptual loss used in single image super-resolution style [39], we employ features 5_4 from the VGG network [44] as a feature perceptual loss:

$$L_{vgg} = \mathbb{E}_{p_{data}} \left[ \| \phi(I_s) - \phi(I_r) \|_2 \right]$$  \hspace{1cm} (9)$$

where $\phi(\cdot)$ is the feature extracted from a pretrained VGG network on ImageNet. Feature perceptual loss is correlated to human perception.

Ultimately, the total loss functions to train our multi-scale dense attention frame interpolation as following. In practice, we set $\lambda_{GAN} = 0.0001$ and $\lambda_{vgg} = 0.001$.

$$L = \lambda_{GAN} (L_F + L_S) + L_f + \lambda_{vgg} L_{vgg}$$  \hspace{1cm} (10)$$

IV. EXPERIMENTS

In this section, we perform extensive experiments to demonstrate the effectiveness of our approach. First, we explain the experiment details. Then, we provide ablation study to evidence the contribution of the proposed method, i.e. 1) multi-scale generator; 2) attention module and 3) sequence adversarial loss. At last, we compare our method with state-of-the-art video frame interpolation methods and evaluate them both qualitatively and quantitatively.

A. EXPERIMENT DETAILS

We can use any available video to train our network, and we do not need any labels about the video. So we use the adobe240-fps video collected by [45] due to real and diverse scenes. We extract three consecutive frames of these videos to form frame triplets. In order to train the model effectively, we discard triplets for which first and third frames is almost identical. Finally, 50k triplets are selected. Among them, randomly selected 783 triplets are used for test and the rest are used for train. When compared with other methods, we also use UCF101 [46] to verify network performance. The triplets are cropped to patches with size of $128 \times 128$ that allows us to avoid patches that contain no useful information, while use original size frames for test. Because using a fully convolution network, our frame interpolation method can be applied to Arbitrary size of input image. Peak-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are widely used objective metrics in video frame interpolation task, and we also use the two metrics for evaluation. Higher values of PSNR and SSIM indicate better results.

To train our model, we use Adam optimization [47] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for generator and discriminator. The initial learning rate is 0.0001 and is decayed every 20 epochs.
B. ABLATION STUDY

We perform ablation study to prove the effectiveness of our contributions: multi-scale generator structure, attention module and sequence adversarial loss. A baseline model is simply based on GANs framework without multi-scale set and attention module in generator. And loss function to train baseline model includes original scale frame adversarial loss, frame reconstruction loss and feature perceptual loss. The adobe240-fps video test dataset are used in these experiments. The results of ablation study are shown in TABLE 1. Our full model denotes that the model use all of proposed three contributions.

As reported in TABLE 1, the baseline model acquires acceptable PSNR and SSIM values. It shows that the GANs framework can use to video frame interpolation task. But the predicted intermediate frame is not good visually, as shown in Figure 5(b). The word in the moving car is blurry. The PSNR and SSIM can be significantly improved by inserting

| Table 1. The results of ablation study. |
|----------------------------------------|
|                                        |
| Baseline                               |
| 29.43                                  |
| 0.883                                  |
| Baseline + Multi-scale                 |
| 31.12                                  |
| 0.911                                  |
| Baseline + Attention                   |
| 30.96                                  |
| 0.910                                  |
| Baseline + Multi-scale + Attention     |
| 32.51                                  |
| 0.923                                  |
| Our full model                         |
| 33.37                                  |
| 0.937                                  |
our improved strategy into the baseline mode. According to two quantitative evaluation metrics, our full model has the best performance than other ablation models. Comparing rows 1 and 2 of TABLE 1, the multi-scale structure is effective because the model can integrate global and local information. Generator only with attention module is a little worse than only with multi-scale structure, this may due to multi-scale loss can be constructed by using multi-scale structure. In addition, multi generator and discriminator can improve training stability of GAN. Using multi-scale structure and attention module simultaneously can get better performance. Comparing the last two rows in TABLE 1, our full model can capture spatial and temporal consistency owing to a sequence discriminator.

The results of models with different combination are shown in Figure 5. Although the baseline model is feasible, it still produces blurry results. Baseline model with multi-scale structure can generate more realistic frame than model with attention. The produced middle frame of our full model is sharper. The latter two models are very similar in visual perception, but the full model has better quantitative performance.

C. COMPARISON WITH STATE-OF-THE-ART METHODS
In order to verify the progressiveness of the proposed method, we compare our network with several state-of-the-art video frame interpolation methods and evaluate them both qualitatively and quantitatively. Comparison methods includes DVF [6], SepConv [5] and Super SloMo [2]. These three methods are representative of various video frame interpolation techniques. DVF and SepConv are directly recent CNN-based methods. And Super SloMo use optical flow estimated by CNN to assist in the generation of intermediate frame. Due to lack of pretrained model, we train our data with official implementation. We make a quantitative and qualitative comparison on UCF101 and adobe240-fps. Hence, the comparison results are persuasive. The PSNR and SSIM of different methods are shown in TABLE 2 and TABLE 3.

As reported in TABLE 2 and TABLE 3, our method is better than other methods. The SSIM value of our model is a little lower than SepConv and Super SloMo on UCF101 test datasets. The PSNR score of our model is best among these approaches. The quantitative comparison proves that our method is effective. Although the SSIM value of our method is not the highest in UCF101 test data, the intermediate frame generated by our model has better visual reality, as shown in Figure 7. Maybe it’s because numerical indicators are not always consistent with human perception. We list the runtime of different methods in TABLE 3. When testing at 640 × 360 resolution with a 1080Ti GPU, our model is more efficient with 0.324s per frame. Other methods take more time to
TABLE 2. The results of comparison study on UCF101.

|          | PSNR(dB) | SSIM |
|----------|----------|------|
| DVF      | 32.26    | 0.918|
| SepConv  | 33.02    | 0.933|
| Super SloMo | 33.17    | 0.933|
| Our model| 33.58    | 0.926|

TABLE 3. The results of comparison study on adobe240 fps.

|          | PSNR(dB) | SSIM | Runtime(s) |
|----------|----------|------|-------------|
| DVF      | 32.40    | 0.920| 1.352       |
| SepConv  | 32.92    | 0.931| 0.416       |
| Super SloMo | 33.09    | 0.928| 0.392       |
| Our model| 33.37    | 0.937| 0.324       |

generate an intermediate frame than ours. And parameters of our generator model are about 3 million. However, the parameters of other methods are more than 10 million. Due to multi-scale generative adversarial structure and attention modules, our model can get better performance with less parameters.

Some frame interpolation results of other methods and our model are shown in Figure 6 and Figure 7. On adobe 240-fps test dataset, our proposed model can produce more realistic intermediate frame, as shown in Figure 6 (e) and (j). In Figure 6 (e), the front tyre of red cars have less blurring. In Figure 6 (j), the heads of the person riding on a bicycle are sharper. The chain and transmission looks more realistic. Figure 7 shows some results of these methods on UCF101 dataset. Column (e) is the result of our method. The edge of hula hoops and paddle are better reconstructed and the texture of the sole is more vivid. The generated intermediate frames of our model have better visual perception and accurate spatial and temporal consistency. Due to the use of multi-scale structure and attention mechanism, our model can focus on moving objects, and combine the global and local features to synthesize the intermediate frame preferably.

V. CONCLUSION

In this paper, we propose a novel multi-scale dense attention generative adversarial network (FI-MSAGAN) for video frame interpolation. First, we establish a generative adversarial framework with a multi-scale loss to produce middle frame in a form of coarse-to-fine. Generators can better combine global and local information by receiving rough results from lower levels. Second, an attention module is embedded in generator so that the network can accurately focus on moving objects and deal with large motion precisely. Third, we adopt a sequence discriminator to judge whether the sequence is true or false. So generators can capture spatial and temporal consistency in frame sequence via feedback information from the sequence discriminator. The ablation study proves the effectiveness of our contributions. We implement comparative experiments with several state-of-the-art video frame interpolation methods and evaluate them both qualitatively and quantitatively. The results show our method outperforms state-of-the-art methods.

It has been shown in recent works that traditional numerical metrics (PSNR, SSIM) are not always consistent with human perception. It can be helpful to evaluate video frame interpolation models by designing new metrics that are more suitable for human cognition. Furthermore, combining GANs with optical flow estimation is also worthy of future research.

REFERENCES

[1] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung, “Phase-based frame interpolation for video,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1410–1418.
[2] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, “Super SloMo: High quality estimation of multiple intermediate frames for video interpolation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 9000–9008.
[3] S.-I. Sekiguchi, Y. Idehara, and K. Asai, “A low-cost video frame-rate up conversion using compressed-domain information,” in Proc. Vis. Comput. Image Process., Jul. 2005, pp. 974–977.
[4] S. Niklaus, L. Mai, and F. Liu, “Video frame interpolation via adaptive convolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2270–2279.
[5] S. Niklaus, L. Mai, and F. Liu, “Video frame interpolation via adaptive separable convolution,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 261–270.
[6] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, “Video frame synthesis using deep voxel flow,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4473–4481.
[7] H. Zhang, R. Wang, and Y. Zhao, “Multi-frame pyramid refinement network for video frame interpolation,” IEEE Access, vol. 7, pp. 130610–130621, 2019.
[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 130610–130621, 2019.
[9] Y. Hong, U. Hwang, J. Yoo, and S. Yoon, “How generative adversarial networks and their variants work: An overview,” ACM Comput. Surv., vol. 52, no. 1, pp. 1:1-1:43, May 2019.
[10] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Efros, “Generative adversarial networks: An overview,” IEEE Signal Process. Mag., vol. 35, no. 1, pp. 53–65, Jan. 2018.
[11] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, “Recent progress on generative adversarial networks (GANs): A survey,” IEEE Access, vol. 7, pp. 36322–36633, 2019.
[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 5967–5976.
[13] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in Proc. 7th Int. Conf. Learn. Represent. (ICLR), May 2019, pp. 1–35.
[14] T. R. Shaham, T. Dekel, and T. Michaeli, “SinGAN: Learning a generative model from a single natural image,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4569–4579.
[15] M. Mathieu, C.ouprie, and Y. Lecun, “Deep multi-scale video prediction beyond mean square error,” in Proc. 4th Int. Conf. Learn. Represent. (ICLR), May 2016, pp. 1–14.
[16] X. Liang, L. Lee, W. Dai, and E. P. Xing, “Dual motion GAN for future-flow embedded video prediction,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 1762–1770.
[17] Y.-H. Kwon and M.-G. Park, “Predicting future frames using retrospective cycle GAN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1811–1820.
[18] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2016, pp. 630–645.
[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[20] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7794–7803.

[21] D. Mahajan, F. C. Huang, W. Matusik, R. Ramamoorthi, and P. Belhumeur, “Moving gradients: A path-based method for plausible image interpolation,” ACM Trans. Graph., vol. 28, no. 3, Aug. 2009, Art. no. 42.

[22] M. Werlberger, T. Pock, M. Unger, and H. Bischof, “Optical flow guided TV-L1 video interpolation and restoration,” in Energy Minimization Methods in Computer Vision and Pattern Recognition. Berlin, Germany: Springer, Jul. 2011, pp. 273–286.

[23] S. Baker, S. Roth, D. Scharstein, M. J. Black, J. P. Lewis, and R. Szeliski, “A database and evaluation methodology for optical flow,” in Proc. IEEE 11th Int. Conf. Comput. Vis., 2007, pp. 1–8.

[24] Z. Yu, H. Li, Z. Wang, Z. Hu, and C. W. Chen, “Multi-level video frame interpolation: Exploiting the interaction among different levels,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 7, pp. 1235–1248, Jul. 2013.

[25] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “FlowNet 2.0: Evolution of optical flow estimation with deep networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1647–1655.

[26] Y. L. Liu, Y. T. Liao, Y. Y. Lin, and Y. Y. Chuang, “Deep video frame interpolation using cyclic frame generation,” in Proc. 33rd AAAI Conf. Artif. Intell., Feb. 2019, pp. 8794–8802.

[27] S. Niklaus and F. Liu, “Context-aware synthesis for video frame interpolation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1701–1710.

[28] H. Li, Y. Yuan, and Q. Wang, “FI-net: A lightweight video frame interpolation network using feature-level flow,” IEEE Access, vol. 7, pp. 118287–118296, 2019.

[29] G. Long, L. Kneip, J. M. Alvarez, H. Li, X. Zhang, and Q. Yu, “Learning image matching by simply watching video,” in Proc. IEEE Int. Conf. Vis. (ICCV), Oct. 2016, pp. 434–440.

[30] H.-E. Ahn, J. Jeong, and J. W. Kim, “A fast 4K video frame interpolation using a hybrid task-based convolutional neural network,” Symmetry, vol. 11, no. 5, p. 619, 2019.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2242–2251.

[32] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 105–114.

[33] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in Proc. Mach. Learn. Res. (ICML), Aug. 2017, pp. 214–223.

[34] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. 6th Int. Conf. Learn. Represent. (ICLR), May 2018, pp. 1–26.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in Proc. Mach. Learn. Res. (ICML), Aug. 2019, pp. 7354–7363.

[36] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, “Least squares generative adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2813–2821.

[37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8798–8807.

[38] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, “Asymmetric GAN for unpaired Image-to-Image translation,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5881–5896, Dec. 2019.

[39] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Vis. (ECCV), Oct. 2016, pp. 694–711.

[40] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6450–6458.

[41] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, “Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2417–2426.

[42] Y. Mojtabi, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, “Unsupervised attention-guided image-to-image translation,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Dec. 2018, pp. 3697–3707.

[43] J. Li, J. Jia, and D. Xu, “Unsupervised representation learning of image-based plant disease with deep convolutional generative adversarial networks,” in Proc. 37th Chin. Control Conf. (CCC), Jul. 2018, pp. 9159–9163.

[44] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. 3th Int. Conf. Learn. Represent. (ICLR), May 2015, pp. 1–14.

[45] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, “Deep video deblurring for hand-held cameras,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 237–246.

[46] K. Soomro, A. Roshan Zamir, and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” 2012, arXiv:1212.0402. [Online]. Available: http://arxiv.org/abs/1212.0402

[47] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. 3th Int. Conf. Learn. Represent. (ICLR), May 2015, pp. 1–12.

[48] S. Wen, W. Liu, Y. Yang, T. Huang, and Z. Zeng, “Generating realistic videos from keyframes with concatenated GANs,” IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 8, pp. 2337–2348, Aug. 2019.

[49] M. Dong, S. Wen, Z. Zeng, Z. Yan, and T. Huang, “Sparse fully convolutional network for face labeling,” Neurocomputing, vol. 331, pp. 465–472, Feb. 2019.

JIAN XIAO received the B.S. degree in communication engineering from Northeast Electric Power University, China, in 2015. He is currently pursuing the Ph.D. degree in information and communication engineering with Harbin Engineering University. His research interests include image processing, generative adversarial networks, video analysis, and computer vision.

XIAOJUN BI received the Ph.D. degree from the College of Information and Communication Engineering, Harbin Engineering University, China, in 2006. She was a Professor with Harbin Engineering University. She is currently a Professor with the Minzu University of China. Her research interests include evolutionary computation, multi-objective optimization, image/video analysis, and deep learning.