Paper:

TEMPORAL LONG-TERM FRAME INTERPOLATION USING U-NET DISCRIMINATOR

Soh Tamura, Shin Kawai, Hajime Nobuhara

University of Tsukuba Graduate School of Systems and Information Engineering Department of Intelligent Interaction Technologies
E-mail: tamura@cmu.iit.tsukuba.ac.jp, kawai@cmu.iit.tsukuba.ac.jp, nobuhara@cmu.iit.tsukuba.ac.jp

[Received 00/00/00; accepted 00/00/00]

Abstract. To improve the accuracy of frame interpolation over long time intervals in video data, we propose a frame interpolation model based on U-Net, which identifies each pixel. The two discriminators associated with existing interpolation methods are each extended to a U-Net configuration to discriminate videos and images for each pixel. By training the generator and discriminator based on the loss values by the U-Net, accurate image generation can be achieved. We performed comparisons on the KTH video dataset, and compared the proposed method with the conventional method. The results indicated that the quantitative scores were almost the same, however the visual evaluation revealed that the proposed method produced more accurate results.

Keywords: Frame Interpolation, Generative Adversarial Network, U-Net, Super-resolution

1. Introduction

In recent years, a technique called frame interpolation has attracted attention in the field of super-resolution[13]. Frame interpolation is a technology that uses image processing to interpolate an intermediate image in a video so that the image can be visually smoothed. Frame interpolation has been studied in the framework of image processing, and in recent years, there has been remarkable development in the interpolation of intermediate images using a machine learning model. In a method proposed by Nikelaus et al, frames before and after interpolation are input to a machine learning model, and through motion estimation based on optical flow and pixel synthesis processing, an intermediate frame is generated [9][10]. The purpose of this method is to interpolate one intermediate frame, in contrast, in the method proposed by Jiang et al, an optical flow inference mechanism and frame interpolation mechanism are constructed using two U-Net architectures to realize multiple interpolation models between continuous images independent of time parameters [3]. Although flow-based machine learning models can effectively interpolate between continuous images (i.e., adjacent images), they are not suitable for tasks with long time intervals, and discontinuous frame interpolation is likely to occur. For example, Crestovao et al. reported a case in which a conventional optical flow-based inference model generated discontinuous and invalid interpolated images when the angle of the content object changed greatly in the preceding and following images [7]. To solve this problem, frame interpolation over long time intervals has recently been proposed as an extension of frame interpolation. Li et al. presented a long-time-range frame interpolation method based on latent representation generators [12]. This method consists of a Generative Adversarial Network (GAN) framework [4].

However, the proposed model has the problem of low image generation accuracy. To solve the problem of scarce loss information being fed back to the conventional model, we propose a frame interpolation method based on the U-Net discriminator [2].

2. Related Research

2.1. Long-term frame interpolation

Existing frame interpolation techniques mainly focus on interpolation between adjacent images, and are not always suitable for frame interpolation over long time intervals, as discontinuous interpolation tends to occur.Several studies have addressed this limitation Chen et al. proposed time frame interpolation over a long time interval using a convolutional neural network (CNN) with two encoders and one decoder to predict frames from both the start and end image directions [11]. This achieves frame interpolation over a long time interval by convolving the start and end image with separate encoders and then combining and decoding the feature maps. Xu et al. used a bi-directional long short-term memory(LSTM) model to allow the model to learn temporal characteristics of a video, thereby enabling the generation of interpolated images with higher accuracy and validity than the previously used CNN model [8]. However, because this model uses the LSTM model as a basis for learning, the complexity of the model increases and learning becomes difficult, Additionally the problem of poor scalability arises depending on the number of interpolated images. To solve this problem, Li et al. proposed a long-range frame interpolation model with improved learning complexity and scalability by constructing a full CNN model based on GAN [12]. The model consists of five CNN models as components

of the network. Additionally, it contains three generators and two discriminators as follows: an image encoder, latent representation generator, video generator, video discriminator, and image discriminator. The image encoder extracts the latent representations of the start image and end image, while the latent representation generator interpolates the latent representations in the time direction. The video generator decodes the interpolated latent representations to the original image size. The video generator then identifies the generated continuous image in the unit of motion and learns by identifying them in the unit of image. The main proposal in [12] is the latent representation generator. It consists of 24 convolutional blocks and interpolates latent representations in stages based on a course-to-fine scheme (1-8 blocks in two sheets, 9-16 blocks in six sheets, 17-24 blocks in 14 sheets). Whether the interpolation image approaches the start or end image is learned at the latent representation level.

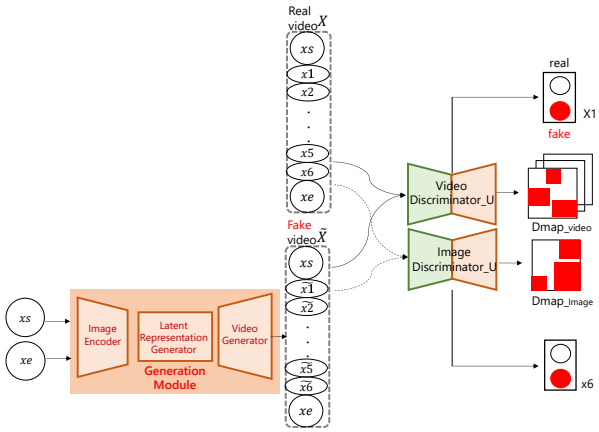

Fig. 1. Proposed framework

2.2. U-Net Model

In recent years, encoder and decoder networks have achieved high accuracy in semantic segmentation tasks, The CNN based on the U-Net model proposed by Ronneberger et al. is widely used, mainly in the field of bioimaging [6]. Specifically, a skip connection exists between the encoder and decoder to obtain a feature map of the same size as the input sample. Then up-sampling is performed. In general, the CNN retains the local feature amount in the deep layer, while the position information of the entire image is lost, whereas in the shallow layer, the overall feature amount is retained. In the context of GAN, Schönfeld et al.proposed a U-Net configuration for the discriminator [2]. It was reported that this enables spatial feedback to the generator and improves the accuracy of the conventional GAN model. Unlike conventional discriminators, which return a single scalar value for a single image, U-Net can provide pixel-by-pixel discrimination results for the entire image. In [8], the authors investigated a single discriminator composed of 2-dimensional(2D) convolutions. In this study, we address a 3-dimensional(3D) convolution for video data and the U-Net configuration of multiple discriminators.

3. Proposed Method

3.1. Extension to U-Net discriminator

The proposed method is based on replacing the conventional latent representation generator with 6-frame interpolation and extending the image discriminator and video discriminator to the U-Net configuration. With regard to the former interpolation number, the conventional model realized 14-frame interpolation; however, we changed it to 6-frame interpolation with the aim of implementing it in a lighter computational environment. Specifically, we eliminated the block corresponding to 16-frame generation of the latent representation generator (17-24 blocks). Therefore, the proposed model has an extended configuration. Moreover, because interpolation between adjacent images is considered to be a sufficient time interval compared with the conventional interpolation between adjacent images, in this study, we used 6-frame interpolation.

We consider the U-Net configuration in interpolation as follows. In the conventional model, the video discriminator is composed of four convolutional layers, whereas in the present model, it is composed of seven 3D convolutional layers and seven transposed 3D convolutional U-Net. This model takes the video of (8, 64, 64, 3) as input, and outputs the identification value of [0, 1], which is scalar. It also outputs the identification map, which distinguishes the entire video pixel-by-pixel. Specifically, this model outputs the scalar identification result after passing through seven 3D convolutional layers, and again passes through seven transposed 3D convolutional layers to output the identification result of the same size as the original image data. In the conventional model, only a single identification value is returned to the video data composed of eight continuous images. In the U-Net configuration of this model, the identification map of the same size as the video size is returned. This improves the authenticity of the entire video and the value of the loss is fed back as the loss value. As a result of trial and error, the mirror-like model, which is closer to the structure of the video generator, has a higher accuracy than the conventional model.

In the conventional model, the image discriminator is composed of eight convolutional layers, and 8 transpose 2D convolutional U-Nets. Particularly, the shortcut (x, c) that exists in the even layer is the residual network(Resnet)-based skip-connection [5]. The model divides the (8, 64, 64, 3) video data into eight images, and inputs (64, 64, 3) images for each image. The model outputs the [0, 1] identification value, which is a scalar, and the identification map, which identifies the entire image for each pixel. In the conventional model, only a single identification result is returned for one image, whereas in the proposed model, the authenticity of the entire image is fed back more clearly as a loss value to return the identification map of the same size as the image size.

3.2. Loss Function

In the conventional model, learning is performed end-to-end by minimizing the following loss function,

$$L_{D_v} = E_{X,\hat{X}}[-logD_v(X) - log(1 - D_v(\hat{X}))], \quad (1)$$

$$L_{D_i} = E_{X,\hat{X}}[\frac{1}{T}\sum_{i=1}^{T}[-logD_i(x_i) - log(1 - D_i(\hat{x}_i))]], \quad (2)$$

$$L_G = E_{X,\hat{X}}[-logD_v(\hat{X}) - \frac{1}{T}\sum_{i=1}^{T} logD_i(\hat{x}_i)], \quad (3)$$

where, D_v and D_i are functions representing the video discriminator and image discriminator, respectively. T represents the number of interpolated images, and in this case, $T = 6$. Additionally, x_i is the interpolated image. After learning the video discriminator and image discriminator at L_{D_v} and L_{D_i}, three generators are learned at L_G: the image encoder, latent representation generator, and video generator.

In the proposed method, the loss function $L_{D^{dec}}$ based on the identification map output from U-Net is added to the conventional loss functions L_{D_v}, L_{D_i}, L_G.

$$L_{D_v^U} = L_{D_v} + L_{D_v^{dec}}, \quad (4)$$

$$L_{D_i^U} = L_{D_i} + L_{D_i^{dec}}, \quad (5)$$

$$L_{D_v^{dec}} = E_{X,\hat{X}}\Big[-\sum_{j,k} log[D_v^{dec}(X)]_{j,k} - \sum_{j,k} log(1 - [D_v^{dec}(\hat{X})]_{j,k})\Big], \quad (6)$$

$$L_{D_i^{dec}} = E_{X,\hat{X}}\Big[\frac{1}{T}\sum_{i=1}^{T}\big[-\sum_{j,k} log[D_i^{dec}(x_i)]_{j,k} - \sum_{j,k} log(1 - [D_i^{dec}(\hat{x}_i)]_{j,k})\big]\Big], \quad (7)$$

Then, the following loss function is minimized where D_v^{dec} is a function that represents the video discriminator of the U-Net configuration and outputs the discriminative map of the video size. D_i^{dec} is a function that represents the image discriminator of the U-Net configuration and outputs the discriminative map of the image size. Subscripts j and k denote the coordinates of the pixels, and the total loss value is calculated. Finally, the loss function of the generator reflects the above equation.

$$L_{G^U} = L_G + L_{G^{dec}}, \quad (8)$$

$$L_{G^{dec}} = E_{X,\hat{X}}\Big[-\sum_{j,k} log[D_v^{dec}(\hat{X})]_{j,k} - \frac{1}{T}\sum_{i=1}^{T}\sum_{j,k} log[D_i^{dec}(\hat{x}_i)]_{j,k}\Big], \quad (9)$$

Adam [12] is used as an optimization algorithm for learning. The parameters are $\beta_1 = 0.5, \beta_2 = 0.999$, and $\varepsilon = 10^{-8}$, and the batch size is 32. Also, based on [10], a constant factor of 0.5 is added to the loss value of the discriminator as in [9]. The learning rate is set to $5 \times 10^{-}5$ and the number of parameter updates is 500,000. The RTX 2080 Ti is used for learning for 7 days.

4. Experimental Results

In this experiment, we performed frame interpolation using the model of the proposed method, and performed quantitative and qualitative evaluation of the generated video images. In this experiment, we used four models of the conventional method and the proposed method, which consisted of U-Net as the model group. In this experiment, we denoted a video discriminator of U-Net as VideoD-U and an image discriminator of U-Net as ImageD-U. In addition, we used KTH [1] as a data set. From this dataset, 15 videos were used as learning data, while the remaining 10 videos were used as test data. Each video was divided into eight frames for learning. Preprocessing of the data generally followed the approach in [12]. In this experiment, the peak signal-to-noise ratio(PSNR) and SSIM [14] were used as image evaluation indices for quantitative evaluation. Quantitative evaluation for the KTH dataset is provided in Table 3.1. Quantitative evaluation of the four models is presented, and quantitative results of the conventional method, ImageD-U, VideoD-U, and proposed method are illustrated in that order. The PSNR had the highest accuracy of 29.9 for the conventional method, while VideoD-U had the highest accuracy of 0.893 for SSIM. This demonstrates that the U-Net model is also effective for the 3D convolutional video discriminator. In contrast, ImageD-U had the lowest accuracy of 24.8 and 0.852 for the PSNR and SSIM, respectively, while proposed method had a slightly lower accuracy of 28.8 and 0.872 for the PSNR and SSIM, respectively. These results indicate that the performance of the proposed method was slightly inferior to the conventional method. Learning appeared to be the bottleneck. The proposed model is a ResNet-based network model with skip connections, and the asymmetric U-Net configuration was likely responsible for the unstable results.The presence of ImageD-U may have reduced the accuracy of learning.

Figures 2-4 present the results for each operation. In the figures, the first row is the ground truth, the second row is the conventional method, the third and fourth rows are ImageD-U, VideoD-U, respectively, and the fifth row is the proposed method. The start image in column 1 and the end image in column 8 are the ground truth,

Table 1. Generation accuracy

Evaluation Index	Conventional Method	ImageD-U	VideoD-U	Proposed Method
PSNR	29.9	24.8	29.7	28.8
SSIM	0.884	0.852	0.893	0.872

while columns 2-7 are the interpolated images. Figure 2 presents the results of the Running test. The conventional method and ImageD-U, generated a slightly distorted head of the person in the video from 2 to 5 frames. There were also frames in which parts of the person disappeared. In the proposed method and VideoD-U, the clothes and legs of the person were generated relatively clearly, and continuity in the video was maintained. In VideoD-U, the problem of the conventional method was avoided by concentrating on the foreground of the person through the pixel-by-pixel feedback by U-Net, and the accuracy of generating each frame was improved. Figure 3 presents the results of the Jogging test. In the conventional method, the feet of the person were blurred in frames 4 and 5. In the proposed method and VideoD-U, the images were generated with high accuracy. Figure 4 presents the results of the Walking test, but the images were generated almost continuously. In the conventional method, the shape of the person's head was lightly distorted, whereas the proposed method, the avoided this problem.

Fig. 2. Running

Fig. 3. Jogging

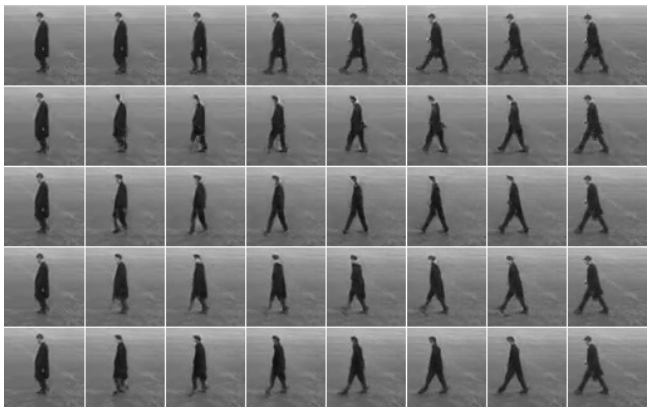

Fig. 4. Walking

5. Conclusion

In this study, we proposed a temporal frame interpolation with the U-Net discriminator to address the problem of low generation accuracy and low loss feedback of the discriminator. To evaluate the effectiveness of the proposed method, we performed quantitative evaluation and qualitative evaluation using the PSNR and SSIM for the KTH video dataset. The results indicated that VideoD-U outperformed conventional methods in quantitative evaluation in terms of the SSIM. The results of visual qualitative evaluation, demonstrated that VideoD-U was generated with temporal continuity. One problem to address in future research is that ImageD-U suffers from unstable learning, resulting in low generation accuracy. The U-Net configuration can be improved by making it closer to the mirror image. It is also necessary to apply this model to a longer time interval. An additional problem is increased computation time. However, this can be improved by changing the video discriminator, which is one of the causes of the increased computation time, to a more efficient configuration.

References:

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach", In International Conference on Pattern Recognition (ICPR), 2004.

[2] E. Schönfeld, B. Schiele, and A. Khoreva, "A U-Net Based Discriminator for Generative Adversarial Networks", In Computer Vision and Pattern Recognition Conference (CVPR), 2020.

[3] H. Jiang, D. Sun, V. Jampani, et al., "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation", In Computer Vision and Pattern Recognition (CVPR), 2018.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets", in Advances in neural information processing systems, pp. 2672–2680, 2014.

[5] K. He, X. Zhang, S. Ren, et al., "Deep residual learning for image recognition", In Computer Vision and Pattern Recognition (CVPR), 2016.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation.", In MICCAI, 2015. 2, 3.

[7] P. Crestovao, H. Nakada, Y. Tanimura, et al., "Generating In-Between Images Through Learned Latent Space Representation Using Variational Autoencoders", Digital Object Identifier, IEEE Xplore, 2020.

[8] Q. Xu, H. Zhang, W. Wang, et al., "Stochastic Dynamics for Video Infilling", Technical report, ArXiv, 2018.

[9] S. Niklaus, L. Mai, and F. liu, "Video Frame Interpolation via Adaptive Convolution", IEEE Conference on Computer Vision and Pattern Recognition, 1:3, 2017.

[10] S. Niklaus, L. Mai, and F. liu, "Video Frame Interpolation via Adaptive Separable Convolution", In International Conference on Computer Vision (ICCV), 2017.

[11] X. Chen, W. Wang, J. Wang, et al., "Long-Term Video Interpolation with Bidirectional Predictive Network", In IEEE Visual Communications and Image Processing (VCIP), 2017.

[12] Y. Li, D. Roblek, and M. Tagliasacchi, "From Here to There: Video Inbetweening Using Direct 3D Convolutions", Arxiv, 2019.

[13] Z. Tu, W. Xie, and D. Zhang, "A survey of variational and CNN-based optical flow techniques", Signal Processing: Image Communication, 72:9–24, 2019.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, et al., "Image quality Assessment: From error visibility to structural similarity", IEEE transactions on image processing, 13(4):600–612, 2004.