Video Frame Rate Up-Conversion via Spatio-Temporal Generative Adversarial Networks

Naomichi Takada and Toshiaki Omori
Department of Electrical and Electronic Engineering, Graduate School of Engineering, Kobe University, Kobe, Japan
Email: omori@eedept.kobe-u.ac.jp

***Abstract*—Video quality has become more important due to the development of information and communication technology. In this study, we propose a spatio-temporal super-resolution method using a Generative Adversarial Network (GAN) in order to achieve a higher frame rate. In recent years, with the development of machine learning technology such as convolutional neural networks, clearer interpolation frame estimation has been realized. Most of the estimation methods use optimization techniques that minimize the mean squared reconstruction error, and the resulting estimates show a high Peak Signal-to-Noise Ratio (PSNR). However, these Mean Squared Error (MSE)-based methods often lack the high-frequency components of the generated frame, resulting in blurry frames. To address this issue, our study adopts GAN that uses spatiotemporal convolution instead of traditional spatial convolution. We propose a method for video frame rate up-conversion with perceptual loss function, which consists of adversarial loss and mean squared loss. This adversarial loss produces a more natural frame using a discriminator network trained to distinguish between the estimated frame and the original frame. We verified the effectiveness of the proposed method using video data containing complex and large motions such as rotational motion and scaling.**

***Index Terms*—video frame interpolation, machine learning, neural networks, deep learning, spatio-temporal data analysis**

I. Introduction

In recent years, with the spread of smartphones, wearable devices, social networking services, the demand for high-quality video images has become extremely large [1]. In general, the quality of a video image is determined by two factors: frame rate and resolution [2]. For high quality video, it is essential to have both high resolution and high frame rate, but it is difficult to achieve both due to data storage, transmission and limitation in imaging device.

Video frame interpolation has been actively studied in the fields of computer vision and video processing [1], [3]. Common conventional frame interpolation methods are based on motion estimation [4]-[6]. The methods based on motion estimation consists of two steps: a motion estimation step to obtain the optical flow between input frames [7], [8], and a generation step using the optical flow to generate intermediate frames. However, in these methods, the accuracy of the final intermediate frame estimation is highly dependent on the accuracy of the optical flow, and in general, it is difficult to generate an accurate optical flow for videos that contain occlusions, large movements, or sudden changes in brightness.

Recently, with the success of machine learning technology, methods that apply deep learning to optical flow estimation [9], image style transformation [10], image correction [11], [12], and image recognition [13], [14] have been proposed. In line with this, Convolutional Neural Network (CNN) based methods for frame interpolation have been proposed [15], [16]. These methods generate an interpolated frame by extracting spatial features from the input frames using a two-dimensional convolutional neural network. Long *et al*. [15] developed a convolutional neural network that interpolates a frame between two input frames by generating the interpolated frame as an intermediate step for estimating the optical flow. A method that considers frame interpolation as a local convolution on two input frames and uses CNN to learn a spatially adaptive convolutional kernel for each pixel has also been proposed, and this method can provide high quality results [16]. However, predicting a kernel for every pixel is computationally expensive and memory consuming, and it cannot deal with movements larger than the kernel [16]. On the other hand, when the number of input frames is set to two, as it is in these methods, the estimation accuracy may decrease for video images containing nonlinear motion. Tanaka and Omori [17] proposed a frame interpolation method for extracting nonlinear motion features using a three-dimensional convolutional neural network based on multiple input frames.

Recent studies have shown that Generative Adversarial Networks (GAN) play an important role in static image super-resolution. Ledig *et al*. [18] proposed a neural network model that realizes the 4× static image super-resolution while maintaining the sharpness of the image. The SRGAN method, a GAN for image Super-Resolution (SR), incorporates the structure of a GAN in addition to the per-pixel error used in conventional methods. The *discriminator* in the GAN discriminates between the true image and the image generated by the *generator*, and these adversarial learnings produce images that are visually pleasing to humans. Following the success of Ledig *et al*. [18], GAN-based image super-resolution methods for static images have been proposed in order to realize higher estimation accuracy [19], [20].

Manuscript received March 17, 2021; revised August 19, 2021.

doi: 10.18178/joig.9.3.87-94

In frame interpolation for videos, conventional methods use Mean Squared Error (MSE) as the loss function for optimization. While MSE-based static image super resolution methods generally show high PSNRs, they produce excessively smooth frames with low perceptual quality for images with complex textures. This is because the MSE-based method uses pixel-by-pixel image differences in order to find an average solution or an average tendency.

In this study, we propose a new frame-interpolation network with four frame inputs using the GAN framework. We use as the loss function of the network a loss function that is the sum of adversarial loss and Mean Squared Error loss (MSE loss). In particular, in order to realize higher estimation accuracy, the original framework of the GAN with perturbation noise is adapted; we train a frame rate up-conversion neural network with a loss function that is the sum of MSE-loss and adversarial loss for output frames obtained from the input frames rather than perturbation noise.

In order to achieve accurate motion estimation from a large number of input frames, a 3D convolutional neural network with spatio-temporal filters is used in this study. In the proposed method, instead of the conventional 2D feature extraction in only the spatial direction, 3D feature extraction in the spatio-temporal dimension is used, which is considered to enable motion feature extraction with higher accuracy.

The structure of this paper is as follows. In Section II, we briefly explain the conventional methods used for frame interpolation. In Section III, we describe in detail the proposed method, which is based on a spatio-temporal super-resolution network using GANs. In Section IV, the effectiveness of the proposed method is verified using the standard benchmark dataset. The concluding remarks are given in Section V.

II. Existing Method

A. Generative Adversarial Networks

The framework of a GAN is shown in Fig. 1. In a GAN, two networks, a *generator* and a *discriminator*, are used for adversarial learning [21]. A GAN is a kind of generative models that can generate non-existent data with realistic characteristics or transform data along the features of existing data by learning features from the data.

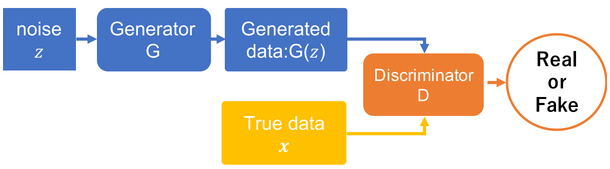


Figure 1. Schematic of a Generative Adversarial Network (GAN). The *generator* generates data from noise z, and the *discriminator* determines whether the data is real or fake.

GANs are attracting attention as a method of unsupervised learning that learns features without being given correct data. Due to the flexibility of their architecture, they can be used in a wide range of domains depending on the idea. Application and theoretical studies are rapidly progressing, the effectiveness of GANs has been demonstrated in the field of spatial super-resolution, and their development is highly anticipated in many fields [18].

The learning process of GANs is expressed by the following equation [21]:

$$\min_G \max_D V(G,D) \\ = \min_G \max_D \ \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))] \quad (1)$$

where D denotes the *discriminator* and G denotes the *generator*. Here, z represents the input noise and $G(z)$ is the data generated by the *generator* and x is true data.

The *discriminator* D tries to determine whether the data generated by the *generator* G is real or fake, and tries to maximize the probability $D(x)$ of labeling it correctly. On the other hand, the *generator* G tries to minimize the probability $\log(1 - D(G(z)))$ that D labels G as fake in order to make D recognize that the generated data is real.

If D is correctly labeled, the value of $D(x)$ becomes large, and $\log D(x)$ also becomes large. Furthermore, if the data generated by G is found to be false, $D(G(z))$ becomes small. As a result, $\log(1 - D(G(z)))$ becomes large and D becomes dominant.

On the other hand, if G can produce data close to the real thing, i.e., if D cannot be labeled correctly, then, the value of $G(z)$ becomes large and $D(G(z))$ also become large. Furthermore, when D cannot be labeled correctly, the value of $D(x)$ becomes smaller and $\log D(x)$ also becomes smaller. As a result, $\log(1 - D(G(z)))$ becomes smaller and G becomes dominant. By repeating the procedure in this method, D and G are updated alternately to deepen the learning process.

B. Spatio-Temporal Convolution

In the proposed network, we apply not only 2D convolution in the spatial dimension, which is often used in conventional frame rate up-conversion methods, but also 3D convolution in the spatio-temporal dimension (i.e., both spatial and temporal dimensions).

We define a new convolution to extract spatio-temporal features. When the feature map of the n-th layer is x^n and the spatio-temporal convolution filter is w^n, the output h^n of the convolution is calculated as follows:

$$h^n_{j,(x,y,z)} = \sum_k \sum_{v,h,t} w^n_{k,j,(v,h,t)} x^n_{k,(x+v,y+h,z+t)} + b_j \quad (2)$$

where x, y, z represents the pixel position of the convolution output, and v, h, t represents the position of the convolution filter. Note that v, h represent indices for convolution in the spatial dimension (vertical and horizontal directions), and t represents an index for convolution in the temporal dimension. In addition, k, j represents the feature map number of the n-th and $(n+1)$-th layer, and b_j represents the bias term.

The main advantage of using 3D spatio-temporal convolution is that it can efficiently extract features from video data with a 3D extent. It is widely known that neighboring pixel values in a static image are likely to be close to each other. In the same way, spatially adjacent pixel values in video data are often close to each other. In other words, the video data has 3D features in the spatio-temporal dimension. In the conventional two-dimensional convolution method, a two-dimensional feature map is generated by convolution in the spatial dimension, and temporal features are lost. The proposed method, on the other hand, generates a 3D feature map by convolution in the spatio-temporal dimension, and thus can effectively utilize the spatio-temporal features.

III. PROPOSED METHOD

In this section, we propose a frame interpolation method using the GAN framework and spatio-temporal convolutional neural network. The most important feature of our method is that it uses the GAN framework for frame interpolation and convolution with spatio-temporal filters. The training method for the proposed neural network is also described.

A. Network Architecture

We show the overall view of the proposed network in Fig. 2. As shown in Fig. 2, the network receives observable frames $\{I_{t-3}, I_{t-1}, I_{t+1}, I_{t+3}\}$ at times $t-3, t-1, t+1, t+3$ and outputs an interpolated frame at time *t*. The network of the proposed method consists of two networks, a *generator* and a *discriminator*.

The *generator* adopts the ResNet [22] structure and consists of three 3D convolutions, nine 2D convolutions and an activation function parametric rectified linear unit (PReLU) [23] after each convolution layer. It takes as input the observable frames $I_t^{in} = \{I_{t-3}, I_{t-1}, I_{t+1}, I_{t+3}\}$ and generates the interpolated frame I_t^{GEN} at time *t* through the 3D convolution and activation functions.

The *discriminator* uses seven 2D convolutions, a leaky rectified linear unit (Leaky ReLU) [24] as the activation function, and a sigmoid function applied to the output layer. The convolution is performed using the frame I_t^{GEN} generated by the *generator* as input, and the output layer outputs a value indicating whether the input frame I_t^{GEN} is a correct frame or a fake frame using the sigmoid function.

The difference between our method and conventional methods is that we use both the *generator* and the *discriminator* to perform adversarial learning. In the conventional methods, only the *generator* is used for learning, and the quality of the generated frame depends on the loss function of the *generator*. For the loss function of the generated frames, MSE-based methods are mainly used [15]-[17]. In the MSE-based method, training proceeds so as to minimize the average error in pixel values between the generated frame and the correct frame. This leads to the problem that the generated frames are excessively smooth [18]. To solve this problem, we propose a method to generate a frame that is closer to the correct image and is superior in terms of quality by adding a loss function that discriminates whether the generated image is the correct image or a fake image using a *discriminator*.

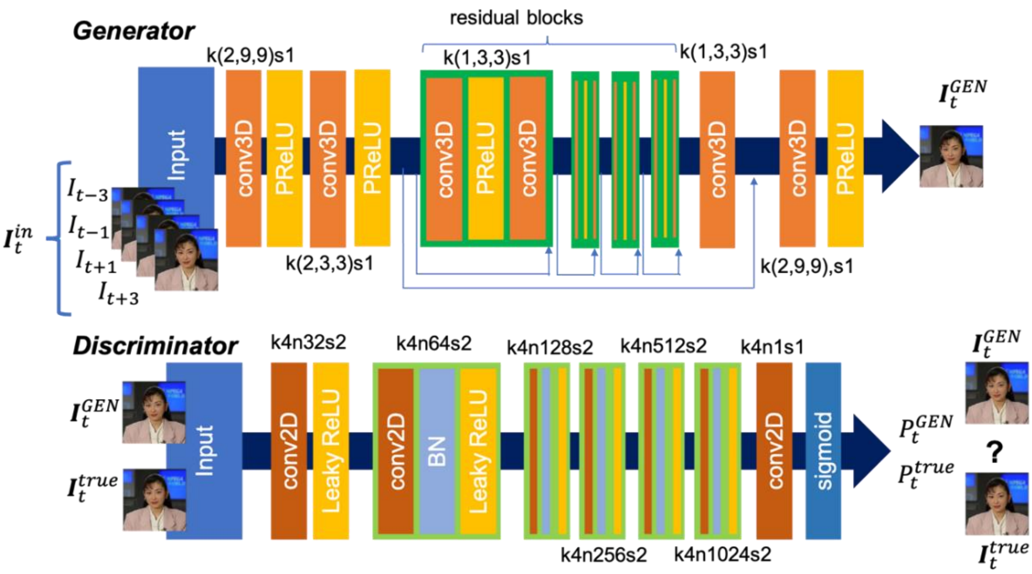


Figure 2. Overall picture of the proposed network consisting of a *generator* and a *discriminator*. The *generator* consists of three 3D convolutions and four residual blocks, and has nine 2D convolutional layers, whereas the *discriminator* has seven 2D convolutional layers. The *generator* generates interpolated frames I_t^{GEN} using the observable input frames $I_t^{in} = \{I_{t-3}, I_{t-1}, I_{t+1}, I_{t+3}\}$. The generated frames I_t^{GEN} and true frames I_t^{true} are provided as inputs to the *discriminator* to discriminate between true and generated data.

B. Training

Here, cost functions for the *generator* and *discriminator* are formulated. We propose a perceptual loss function for video frame rate up-conversion consisting of adversarial loss and mean squared loss. This adversarial loss produces a more natural frame using a *discriminator* network trained to distinguish between the estimated frame and the original frame.

For the *generator*, we used the sum of the mean squared error l_{MSE} between the estimate and the ground-truth and the adversarial loss l_{GAN} as follows:

$$l_{GEN} = l_{MSE} + \lambda l_{GAN} \tag{3}$$

where l_{MSE} and l_{GAN} are expressed as follows:

$$l_{MSE} = \sum_{t \in \mathcal{T}} \| I_t^{true} - I_t^{GEN} \|_2^2 \tag{4}$$

$$l_{GAN} = -\sum_{t \in \mathcal{T}} \log D\left(G\left(I_t^{in}\right)\right) \tag{5}$$

here $\mathcal{T}$ is a set of unobservable times and I_t^{true} is a true frame at time *t*. λ is a hyperparameter in the proposed method. A larger value of λ results in a loss function that is more sensitive to adversarial losses. On the other hand, when the value of λ becomes small, the loss function approaches the mean squared error. The generative loss l_{GAN} in (5) is based on the probability of the *discriminator* $D\left(G\left(I_t^{in}\right)\right)$ for all training samples, where $D\left(G\left(I_t^{in}\right)\right)$is the probability that the generating frame $G(I_t^{in})$ is a natural intermediate frame. To obtain a better gradient behavior, instead of $\log[1 - D\left(G\left(I_t^{in}\right)\right)]$, we minimize $-\log D\left(G\left(I_t^{in}\right)\right)$ [21].

For the *discriminator*, we used the binary cross entropy loss shown in the following equation:

$$l_{DIS} = -\sum_{t \in \mathcal{T}} (\log P_t^{true} + \log(1 - P_t^{GEN})) \tag{6}$$

where P_t^{true} is the output value when the true frame is taken as the input of the *discriminator*, and P_t^{GEN} is the output value when the frame generated by the *generator* is input to the *discriminator*.

During training, both true frame I_t^{true} and the frame generated by the *generator* I_t^{GEN} are provided as an input to the *discriminator* separately. The label for the correct frame is set to be one, and the label for the frame generated from the *generator* is set to be zero. We provide respective labels for output values obtained from the *discriminator* through convolutions and the sigmoid function, and perform training of the network by minimizing the loss function l_{DIS}.

By alternately training the two networks of the *generator* and *discriminator*, our *generator* generates solutions that exist in the natural image diversity by trying to fool the *discriminator*.

IV. Experiment

In this section, we evaluate the effectiveness of the proposed frame interpolation method by means of visual and quantitative comparison.

A. Experimental Settings

To demonstrate the effectiveness of the proposed method, we conducted an experiment using standard benchmark data.

The initial values of the network were determined randomly, and ADAM [25] was used as the network optimizer. The parameters of ADAM were $\alpha = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The size of the mini-batch was set to 10. In this experiment, the filter size in the spatial direction of each convolutional layer of the input and output layers of the *generator* was set to 9×9 pixels, and the filter size of the other *generators* and *discriminators* was set to 3×3. For the test data, we estimated interpolated frames from low-frame-rate video on five different video datasets with a resolution of 352×288 pixels for the standard benchmark data. The training data consisted of five consecutive frames with a resolution of 448×256 pixels in the dataset available on Vimeo90K [26]. To handle large movements, we used 40000 sets with large movements out of the total 64612 datasets.

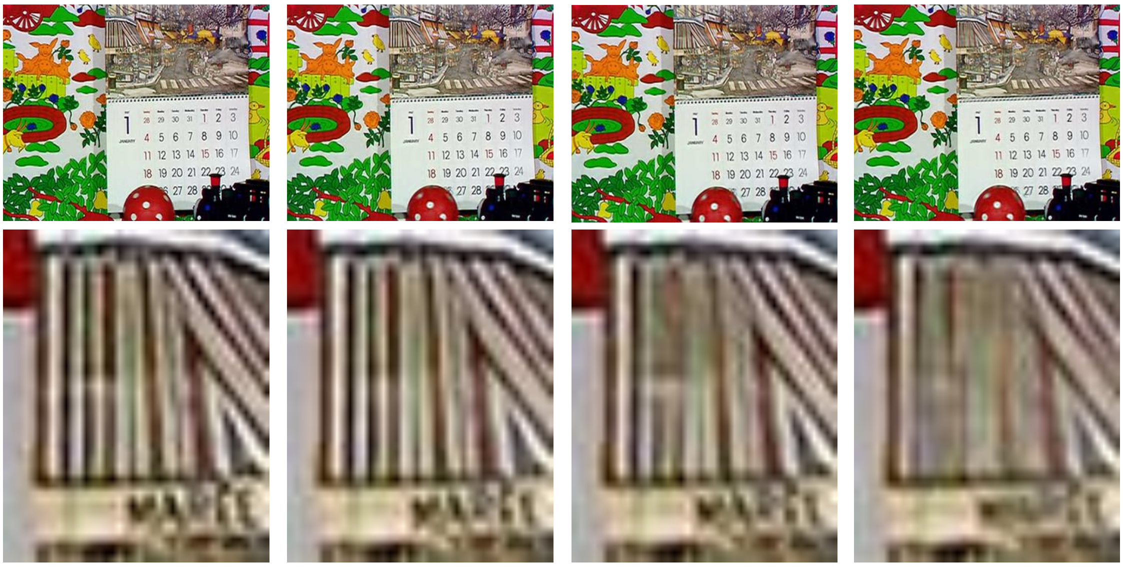
The hyperparameter λ of the loss function (3) was varied for {0.0001,0.001,0.01} in order to show the effect of the adversarial network. Note that a large value of λ corresponds to the case where the effect of adversarial loss on the frame interpolation is expected to be large.

In the experiment, we compared the proposed method with the CNN methods based on the conventional MSE based loss that were proposed by Long *et al.* [15] and Tanaka and Omori [17]. These methods require about the same amount of computation. *Generator only* is a network trained using only the *generator* and the loss function (4) without any *discriminator* in the proposed method. The proposed method and existing methods were trained on the same dataset with the same amount of training. In addition to the Peak Signal-to-Noise Ratio (PSNR), which is the most common image quality metric, we used as an evaluation metric the structural similarity (SSIM), which matches human appearance more closely. As described by Ledig *et al.* [18] and Blau and Michaeli [27], however, the PSNR is a metric based on the MSE and does not necessarily represent human perceptual quality. In this paper, we have shown that our method can produce frames that are more natural to the human eye, even though the PSNR value is low.

B. Visual Comparison

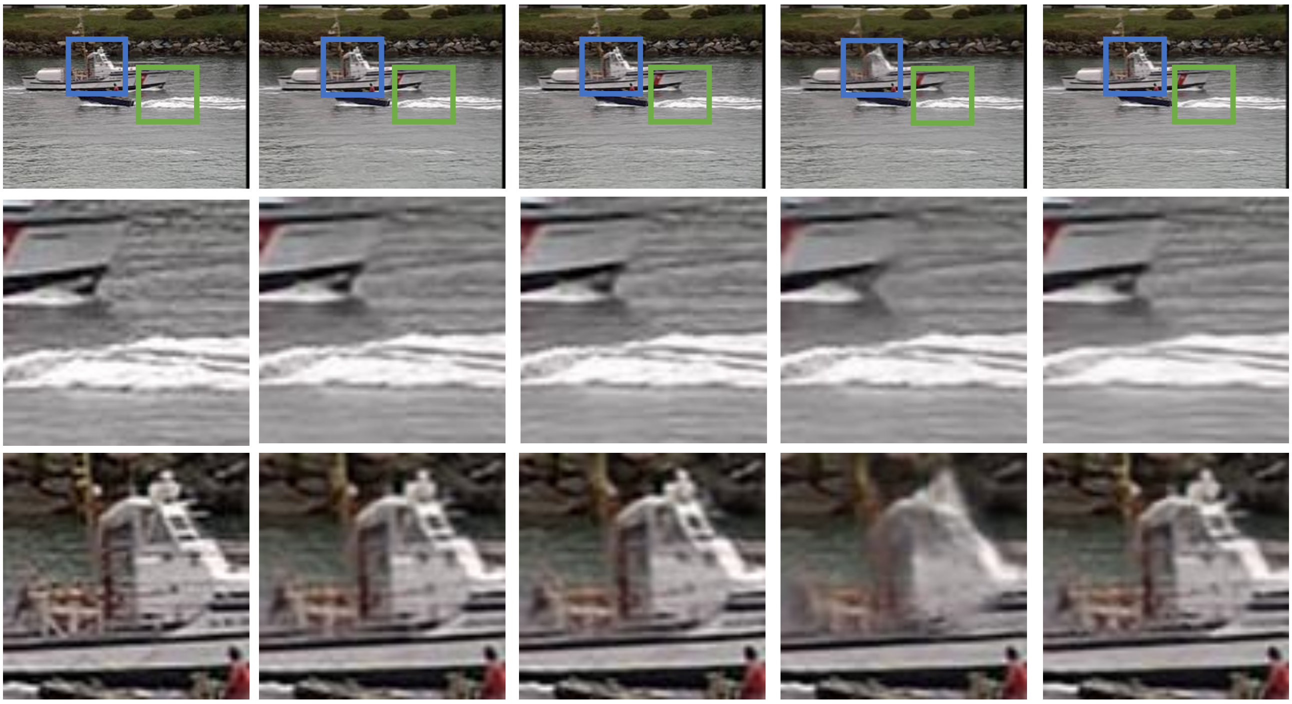
We visually compare the estimation results of the interpolated frames. Fig. 3 shows the results of interpolating frames by changing the hyperparameter λ of the perceptual loss of the proposed method for $\lambda = 0.0001$, 0.001 and 0.01. As we can see in Fig. 3, the stripes on the roof are clearly reproduced for large value of the hyperparameter λ. Namely, we found in Fig. 3 that when the ratio of adversarial loss in the perceptual loss becomes larger, the interpolated frame becomes more natural than the averaged frame. This result indicates that the adversarial generation network can be used for frame interpolation that reproduces edges more clearly. Therefore, the hyperparameter λ of the perceptual loss of the proposed method is set to 0.01.

Fig. 4 shows an example of the frame interpolation results for *coastguard*, which is a video of two boats crossing the coast. We focus on the waves behind the smaller boat and the hull structure of the larger boat. As shown in Fig. 5, the waves indicated by the blue arrows are well reproduced by the proposed method and the Tanaka and Omori's method [17]. The black part of the texture of the wave pointed by the orange arrow is well reproduced by the proposed method and the *generator* alone, but not by Tanaka and Omori [17] and Long *et al*. [15], resulting in an overall white wave. In addition, as shown in Fig. 6, the white bar pointed by the blue arrow is reproduced well by the proposed method, while the shape of the bar is not reproduced straight by the other methods. As for the window of the ship pointed by the orange arrow, the proposed method reproduces it well, but the method using only the *generator* produces a round shape, while the other methods produce a blur. The reason for these blurred interpolated frames and unclear structures is that the MSE based optimization method estimates the frame using the average of pixel values, which produces an excessively smooth frame. On the other hand, the frame interpolation by the proposed method using adversarial learning shows high accuracy in frame interpolation of videos with complex structures and textures.



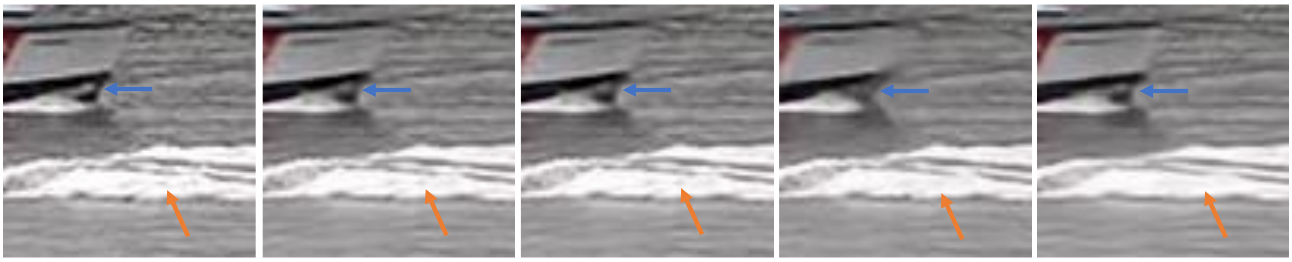
(a) Ground-truth (b) $\lambda = 0.01$ (c) $\lambda = 0.001$ (d) $\lambda = 0.0001$

Figure 3. Frame interpolation results when the hyperparameter λ of the perceptual loss of the proposed method is varied for 0.0001, 0.001, and 0.01. When the hyperparameter λ is increased, the stripes on the roof are clearly reproduced.



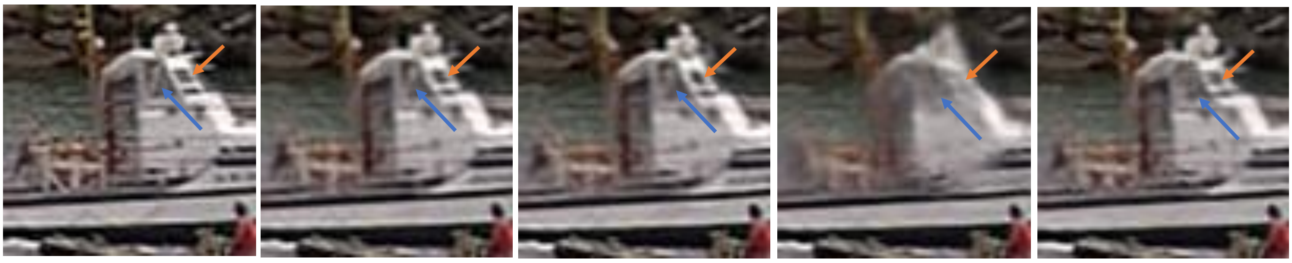
(a) Ground truth (b) Proposed (c) Generator only (d) Tanaka and Omori (e) Long *et al.*

Figure 4. Evaluation of visual quality by using various frame interpolation methods. Evaluation for a video *coastguard* with complex structures and textures. In the case of subfigure (c) [*Generator only*], only the *generator* is trained using only the MSE loss expressed in (4), without using the *discriminator* in the proposed method.



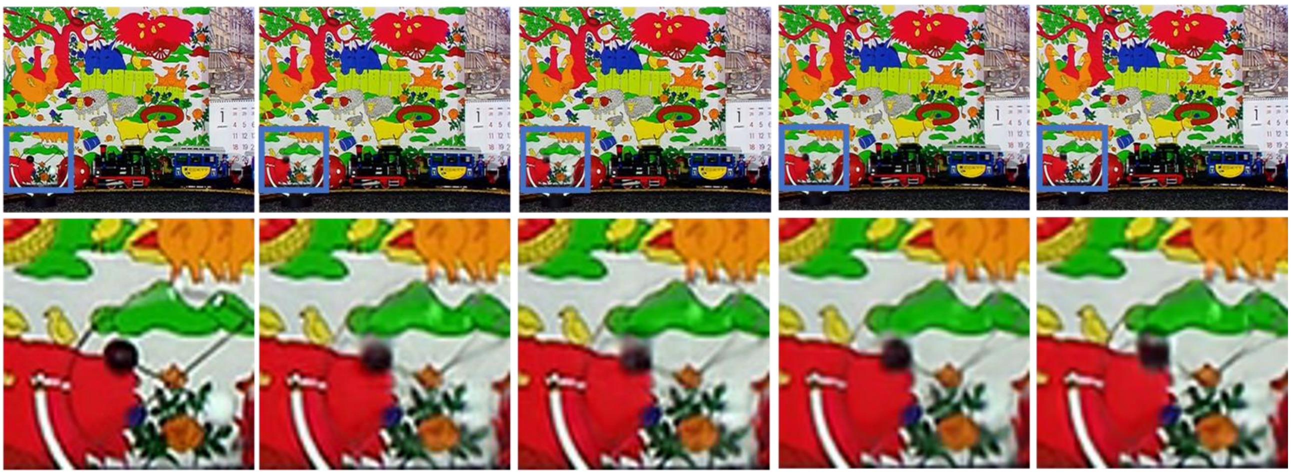
(a) Ground truth (b) Proposed (c) Generator only (d) Tanaka and Omori (e) Long *et al.*

Figure 5. Enlarged view of the center row in Fig. 4. The waves indicated by the blue arrows are well reproduced only by the proposed method and the *generator only*, while the waves are not reproduced by the method of Tanaka and Omori [11]. The black part of the texture of the wave indicated by the orange arrow is well reproduced by the proposed method and the *generator*, but not by the methods of Tanaka and Omori and Long *et al*. resulting in a white wave overall.



(a) Ground truth (b) Proposed (c) Generator only (d) Tanaka and Omori (e) Long *et al.*

Figure 6. Enlarged view of the lowest row in Fig. 4. The white bar pointed by the blue arrow is well reproduced by the proposed method, while the shape of the bar is not reproduced straight by the other methods. As for the window of the ship pointed by the orange arrow, the proposed method reproduces it well, while the *generator only* method reproduces it in a round shape, and the other methods cause blurring.



(a) Ground truth (b) Proposed (c) Generator only (d) Tanaka and Omori (e) Long *et al.*

Figure 7. Evaluating the visual quality of various frame interpolation methods. Evaluation for a video *mobile* with a large rotational motion.

Next, Fig. 7 shows an example of the frame interpolation results for *mobile*. The *mobile* contains various complex motions, such as a calendar moving up and down and a purple sphere rotating. We focus on the purple sphere in rotational motion. The movement of the purple sphere is large and includes nonlinear motion. With Long *et al*. [15] method and *generator only* method, there appear to be two purple spheres. This result suggests that the conventional motion estimation method cannot deal with nonlinear motion like rotational motion. In contrast, the proposed method estimates a clear pattern. This indicates that the proposed method is effective for nonlinear motion.

C. *Quantitative Evaluation*

We quantitively evaluated the estimation accuracy of the interpolated frames by using their PSNR and SSIM values. Those shown in Tables I and II are the averages of PSNR and SSIM values of all interpolated frames generated for the five types of videos in the standard dataset. As can be seen from Table I, the proposed method shows higher estimation accuracy compared to the conventional method for most of the videos used in the experiments. On the other hand, the PSNR value of the proposed method was lower than that of the case where the proposed method was trained using only a *generator* without a *discriminator*. This is because the PSNR is an evaluation metric calculated using the average of the pixel errors, and in the *generator only* training, only the MSE of the pixel values is used as the loss function. In addition, it was shown that the proposed method is particularly effective for *Mobile*, which contain nonlinear and large motions and complex textures. It is difficult to interpolate complex structures by using conventional frame interpolation, but the proposed method with adversarial generation network can accurately estimate such complex structures.

TABLE I. COMPARISON OF PSNR [DB] FOR STANDARD BENCHMARK DATA

video	Proposed	Proposed (Generator only)	Tanaka and Omori	Long *et al.*
Coastguard	30.351	**30.352**	27.820	29.333
Foreman	31.957	**32.142**	30.543	31.310
Ice	31.762	31.765	29.671	**32.236**
Mobile	**28.912**	28.745	26.607	26.682
News	34.355	**35.009**	34.603	33.219
Average	31.375	**31.603**	29.849	30.560

TABLE II. COMPARISON OF SSIM FOR STANDARD BENCHMARK DATA

video	Proposed	Proposed (Generator only)	Tanaka and Omori	Long *et al.*
Coastguard	**0.938**	0.936	0.892	0.930
Foreman	0.941	0.942	0.926	**0.948**
Ice	0.972	0.972	0.963	**0.977**
Mobile	**0.972**	0.969	0.946	0.950
News	0.985	**0.986**	0.985	0.983
Average	**0.962**	0.961	0.942	0.958

As shown in Table II, the obtained SSIM values indicate that the proposed method shows a better frame interpolation performance than the conventional methods in the case of complex nonlinear motions. In other words, the generative adversarial network in the proposed method can produce more natural frame interpolation as seen by humans.

V. CONCLUDING REMARKS

In this paper, we proposed a frame interpolation method using generative adversarial networks as a frame rate up-conversion method for videos with nonlinear and large motion and complex textures. In the conventional method, the mean squared error is used for optimization, which results in excessively smooth frames. In addition, since the interpolated frame was estimated using a two-dimensional convolutional neural network from two input frames, it could not deal with non-linear motion. In the proposed method, a three-dimensional convolutional neural network with a spatio-temporal filter is used to estimate the interpolation frame.

In order to verify the effectiveness of the proposed method, we conducted experiments using video images with complex textures, edges, and nonlinear motions such as rotational motion and human motion. As a result, we found that the proposed method produced interpolated frames with better visual quality than the conventional method or the *generator only* method in regions with complex textures and nonlinear motions. In addition, numerical evaluation by PSNR values was performed. Moreover, the proposed method outperformed the conventional method for many videos in the numerical evaluation by PSNR values. In the numerical evaluation by the SSIM value, the accuracy was higher than that of the method using only the *generator* without the *discriminator* and using only the mean squared error. This indicates that the proposed method with the generative adversarial network performs frame interpolation better than the method with only a *generator*.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Naomichi Takada and Toshiaki Omori performed research; Naomichi Takada and Toshiaki Omori analyzed the data; Naomichi Takada and Toshiaki Omori wrote the paper; all authors had approved the final version.

ACKNOWLEDGMENT

This work is partially supported by Grants-in-Aid for Scientific Research for Innovative Areas “Initiative for High-Dimensional Data driven Science through Deepening of Sparse Modeling” [JSPS KAKENHI Grant No. JP25120010] and for Scientific Research [JSPS KAKENHI Grant No. JP16K00330], and a Fund for the Promotion of Joint International Research (Fostering Joint International Research [JSPS KAKENHI Grant No. JP15KK0010] from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency, Japan.

REFERENCES

[1] R. Szeliski, *Computer Vision*: *Algorithms and Applications*, Springer, 2000.

[2] G. J. Sullivan and T. Wiegand, “Rate-Distortion optimization for video compression,” *IEEE Signal Processing Magazine*, vol. 15, pp. 74-90, 1998.

[3] Z. Xiong, X. Sun, and F. Wu, “Robust web image/video super-resolution,” *IEEE Trans. Image Process.*, vol. 19, pp. 2017-2028, 2010.

[4] B. T. Choi, S. H. Lee, and S. J. Ko, “New frame rate up-conversion using bi-directional motion estimation,” *IEEE Trans. Consumer Electronics*, vol. 46, pp. 603-609, 2000.

[5] M. T. Orchard and G. J. Sullivan, “Overlapped block motion compensation: An estimation-theoretic approach,” *IEEE Trans. Image Process.*, vol. 3, pp. 693-899, 1994.

[6] B. D. Choi, J. W. Han, C. S. Kim, and S. J. Ko, “Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation,” *IEEE Trans. Circuits and Systems for Video Tech*., vol. 17, pp. 407-416, 2007.

[7] J. Jain and A. Jain, “Displacement measurement and its application in interframe image coding,” *IEEE Trans. Communications*, vol. 29, pp. 1799-1808, 1981.

[8] G. D. Haan, P. W. Biezen, H. Huijgen, and O. A. Ojo, “True-Motion estimation with 3-d recursive search block matching,” *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 3, pp. 368-379, 1993.

[9] A. Dosovitskiy, *et al*., “Flownet: Learning optical flow with convolutional networks,” in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 2758-2766.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” arXiv preprint arXiv:1508.06576, 2015.

[11] K. Yu, C. Dong, C. C. Loy, and X. Tang, "Deep convolution networks for compression artifacts reduction," arXiv preprint arXiv:1608.02778, 2016.

[12] M. Tassano, J. Delon, and T. Veit, "Dvdnet: A fast network for deep video denoising," in *Proc. IEEE International Conference on Image Processing*, 2019, pp. 1805-1809.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1904-1916, 2015.

[14] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 2961-2969.

[15] G. Long, L. Kneip, J. M. Alvarez, H. Li, X. Zhang, and Q. Yu, "Learning image matching by simply watching video," in *Proc. European Conference on Computer Vision*, 2016, pp. 434-450.

[16] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 261-270.

[17] Y. Tanaka and T. Omori, "Spatio-temporal convolutional neural network for frame rate up-conversion," in *Proc. 3rd International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, 2019, pp. 35-39.

[18] C. Ledig, *et al.*, "Photo-Realistic single image super-resolution using a generative adversarial network," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681-4690.

[19] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a GAN to learn how to do image degradation first," in *Proc. European Conference on Computer Vision*, 2018, pp. 185-200.

[20] C. You, *et al.*, "CT super-resolutiongan constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE)," *IEEE Trans. Medical Imaging*, vol. 39, no. 1, pp. 188-203, 2020.

[21] I. Goodfellow, *et al.*, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672-2680, 2014.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE International Conference on Computer Vision*, 2016, pp. 1026-1034.

[24] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic model," in *Proc. International Conference on Machine Learning*, 2013, pp. 1-6.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[26] 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, pp. 1106-1125, 2019.

[27] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6228-6237.

Naomichi Takada was born in 1996. He received his B.E. degree from Kobe University in 2019. Now he is pursuing his M.E. in Graduate School of Engineering, Kobe University. His research interests include machine learning theory and its applications.

Toshiaki Omori received his B.S. degree in physics from University of Tsukuba in 1999, and his Ph.D. degree in information science from Tohoku University in 2004. He was a predoctoral research fellow of Japan Society for the Promotion of Science (JSPS) from 2003 to 2004, a postdoctoral researcher at Japan Science and Technology Agency (JST) from 2004 to 2006, and a postdoctoral research fellow of JSPS from 2006 to 2008. He was a visiting researcher at University of Arizona in 2007. He became a research assistant professor and an assistant professor at the University of Tokyo in 2008. He is currently an associate professor at the Graduate School of Engineering, Kobe University. His research interests include machine learning theory and its applications, data-driven science, probabilistic information processing, and computational neuroscience.