

An Analysis of Solutions to the Frame Consistency Problem in Video style Transfer based on GAN

Yuhang Hu ^{1,*}, Zeyu Zhou ²

¹ Department of International Engineering, Xi'an University of Technology, Xi'an, China

² College of Electronic information Engineering, Hubei University, Wuhan, China

* Corresponding author: 3212231061@stu.xaut.edu.cn

Abstract. Video style transfer has emerged as a significant area within the realm of neural style transfer, holding promise for diverse applications. In this context, GANs have gained traction for their advantages in addressing temporal consistency concerns. This paper explores various strategies to manage frame consistency within GAN-based video-style conversion. It delves into techniques such as Recurrent Neural Networks (RNNs), 3D convolutions, inter-frame continuity in the discriminator, and time loss functions. The research highlights key studies that employ RNNs to complement GAN-based frameworks for improved temporal consistency. The study also evaluates a range of other approaches, such as introducing inter-frame continuity in the discriminator and applying time loss functions to minimize visual discrepancies. Through this analysis, the paper contributes insights into the evolving landscape of video style transfer techniques, guiding researchers towards effective strategies for achieving frame consistency and advancing the field.

Keywords: Video style transfer, Generative Adversarial Networks (GANs), temporal consistency.

1. Introduction

Video style transfer, an essential subset of neural style transfer, originates from the field of image style transfer research. It holds broad application prospects in the film and entertainment industries. Over recent years, we have witnessed numerous researchers achieve innovative results in the study of video style transfer, introducing several representative frameworks. However, there remain unresolved issues in video style transfer, such as low temporal consistency during the transformation process and frame jumping issues during continuous stylization. Furthermore, research findings from the image style transfer domain cannot be directly applied to video style transfer studies. Most solutions to these problems are related to Generative Adversarial Networks (GANs). Methods based on GANs that address frame consistency require further organization and analysis.

Consequently, this article compiles and summarizes four GAN-based techniques to handle frame consistency issues: the use of Recurrent Neural Networks (RNN), the application of 3D convolutions, introducing frame continuity in the discriminator, and the implementation of a temporal loss function. By comparing the strengths and weaknesses of these four approaches and analyzing representative research papers on computation speed, computational load, and processing effects, this study identifies the most effective current method based on GANs for addressing frame consistency, specifically using the RNN approach. This will aid researchers in the relevant fields to more precisely determine their research direction, thereby further propelling the development of this domain.

2. Techniques for managing frame consistency utilising generative adversarial networks

The Generative Adversarial Network (GAN)-based video-style conversion is increasingly popular due to its significant advantages. To address the central issue of temporal consistency, various researchers in related fields have proposed diverse improvement methods using generative adversarial network (GAN) as the main framework. After reviewing several authoritative literatures, we summarized the following four methods to deal with frame consistency: the use of Recursive Neural

Network (RNN), the use of 3D convolution, the introduction of inter-frame continuity in the discriminator, and the use of a time loss function.

2.1. Using Recurrent Neural Networks (RNN)

The Recurrent Neural Network (RNN) is a neural network model capable of processing sequential data. It accepts sequences as input (e.g., data points in a time series) and processes them through a recurrent structure that captures the sequence's temporal information. The RNN uses a hidden state to store the model's memories of the sequence, enabling it to hold the memory of the entire sequence. To solve the frame consistency problem, RNN is applied to learn frame-to-frame correlations. With this objective, the RNN processes every frame of a video as input, transmitting the previous time step's hidden state to the current one, and acquiring the correlation relationship between the frames. RNN is applied to solve the frame consistency problem by taking each frame of the video as input and passing the hidden state of the previous time step to the current time step. This way, RNN can learn the correlation relationship between frames to generate the output sequence with continuity and maintain consistency between neighbouring frames during video processing.

All these following articles [1-7] employ Recurrent Neural Networks (RNN) to complement the Generative Adversarial Network (GAN)-based network framework. The goal is to strengthen temporal coherence, create alternative temporal effects, and visually transform the video style.

Hugo Thimonier et al. propose a post-processing model to solve the temporal consistency issue [1]. The model is agnostic to applications like video wind transfer or image processing with GANs, and adopts a recurrent neural network. The model undergoes training through a ping-pong procedure, with the matching loss, presenting a novel technique for GAN video creation that safeguards perceptual loss. The method was assessed using the DAVIS and videvo.net datasets and was discovered to achieve better preservation of the video's overall style, as illustrated in Fig. 1, in comparison to current state-of-the-art post-processing techniques [2]. This comparison demonstrates more effective results in terms of flicker elimination.



Figure 1. A comparison between the post-processing results of the video frames using the end-to-end approach and the post-processing model.

The method eliminates optical flow estimation during inference and effectively avoids newly appearing losses caused by stylistic deterioration present throughout the video. However, it involves processing the original, unconverted video, which results in increased complexity of the task.

Also, in video frame segmentation, many existing modern methods have achieved remarkable results, but accurate foreground object segmentation in the presence of infrequent foreground object motion, sudden background illumination changes, bad weather and dynamic backgrounds is a challenging task. Prashant W. Patil et al. proposed an end-to-end GAN, RMS-GAN, with recursive techniques for Mobile Object Segmentation (MOS) [3]. The proposed RMS-GAN employs recursive techniques based on residual and weight sharing to combine foreground probabilistic knowledge effectively for accurate segmentation. Cyclic residual architecture is proposed in Generator-1 (Gen-1) to obtain a priori foreground probability map (FPM). Furthermore, in Generator-2 (Gen-2), a looping technique is employed to improve the spatial consistency of the estimated FPM by sharing various level features and outputs of the previous frame to estimate the foreground probability map

of the current frame. Temporal consistency during loop training is ensured by sharing network parameters at different time moments (Figure 2).

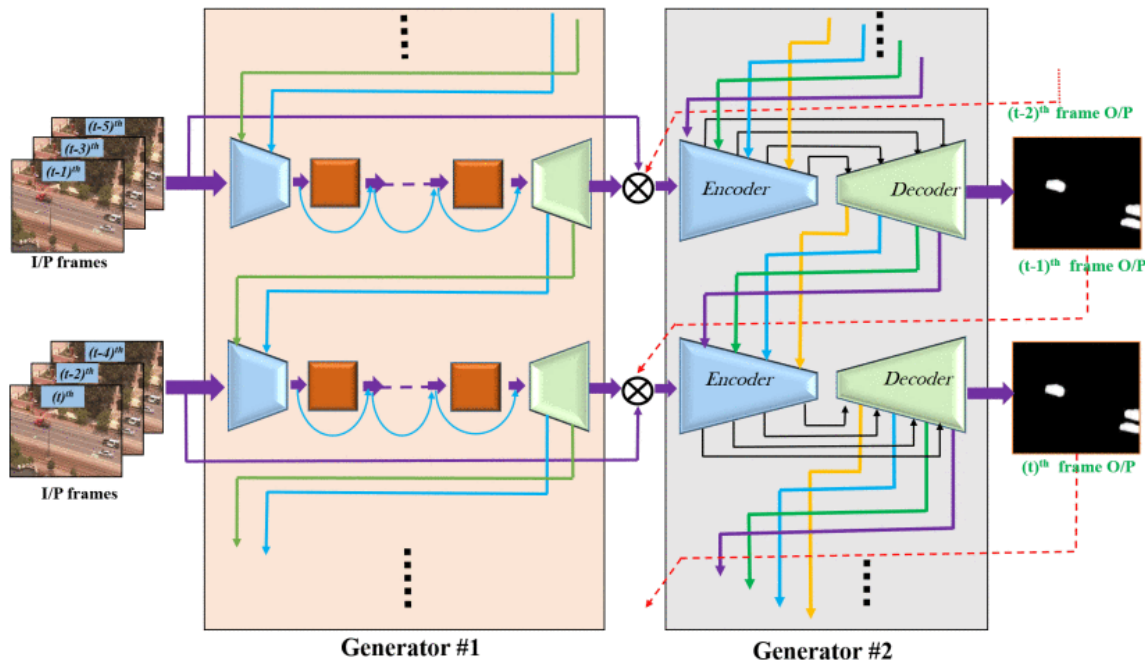


Figure 2. Generalised block diagram of the proposed MOSRMS-GAN framework

Lingjie Xiang et al. suggested an innovative end-to-end GAN structure (CRVC-GAN), which is dedicated to solving the problems of temporal consistency and colouring quality in the field of video colouring [4]. To tackle the issue of temporal consistency, the network structure of Recurrent Neural Network (RNN) is employed, as its shape does not change with the input length and computations consider the historical information to maintain correlation between consecutive frames. According to experimental results, CRVC-GAN attains a superior balance between temporal consistency and colour quality, displaying excellent temporal consistency.

In an article titled "Video Enhancement with Task-Oriented Flow" [5], the authors suggest the use of task-oriented flow, rather than optical flow, to align the frames in a video sequence. According to them, this technique avoids inaccuracies that could arise from optical flow computation in complex situations. The authors combine the RNN with other networks to utilize the RNN's capacity to remember historical information in the sequence. This information is used to reduce noise while maintaining frame consistency during video processing. Experiments have revealed that RNN recurrent neural networks offer substantial advantages. However, these networks require the sequential processing of input data according to the time sequence, making simultaneous computations challenging and resulting in a relatively slow training and inference process. In addition, consistent frame maintenance may not be feasible for very long video sequences.

2.2. Using 3D Convolution

Many researchers prefer using Generative Adversarial Network (GAN) based processing combined with 3D convolutional methods to solve the issue of temporal consistency in video style migration. 3D convolution is different from 2D convolution which only addresses the height and width of the image when dealing with video style migration. It can also address the temporal dimension, i.e., the depth of the video. This grants 3D convolution the ability to naturally take the temporal dimension into account when performing style migration, which leads to better handling of the video dynamics and reduces the discontinuity from frame to frame.

The paper titled "Spatio-temporal video autoencoder with differentiable memory" [8] presents an autoencoder model that is specifically designed to process spatio-temporal data which comprises spatial and temporal dimensions, such as video. As 3D convolution is a commonly used convolution operation for video processing and analysis, it can naturally handle three-dimensional data. This is

why it is used in the nested memory module of convolutional LSTM units in the paper, as illustrated in Figure 3.

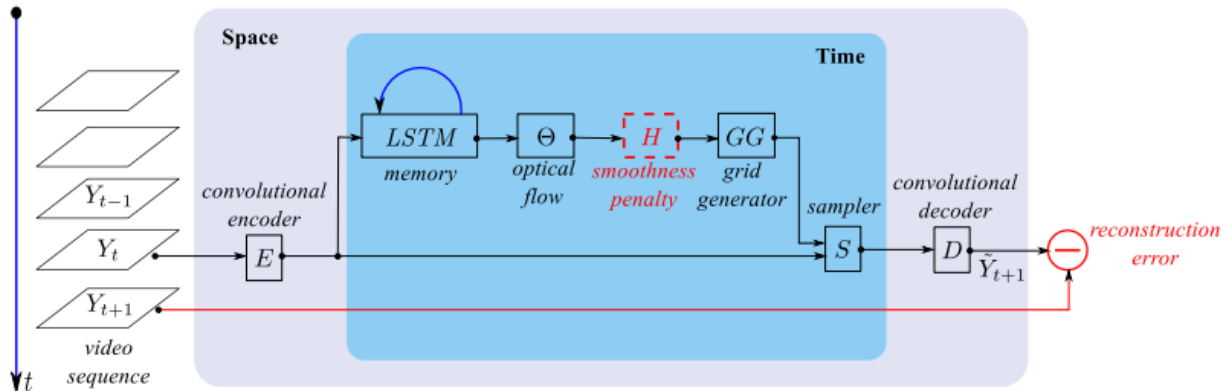


Figure 3. Spatio-temporal video autoencoder

Yan et al. have proposed a novel approach for detecting temporal irregularity in videos by applying a discriminative framework based on 3D convolutional generative adversarial networks (3D-GANs) [9]. The method involves creating a discriminator for 3D-GAN, which identifies temporal irregularities in videos that may indicate erroneous data. The proposed method only uses regular videos in the training phase to detect anomalies based on the bias estimated by the 3D-GAN discriminator. The designers use regular videos as input and construct a 3D-GAN to learn the distribution of regular videos during the training phase. The well-trained discriminator of the network can detect temporal regularity and irregularity as the test data includes irregular videos or spurious data having a different distribution than regular videos or real data.

Experiments conducted on a dataset for detecting video anomalies demonstrate that 3D-GANs are effective and surpass 2D-GANs in identifying temporal irregularities. Although this method gives good results, it is computationally expensive due to the requirement of processing the entire video sequence.

Additionally, S. Aigner and M. Körner proposed the advanced Future GAN model, which is an encoder-decoder GAN model that predicts future frames of a video sequence based on the sequence of past frames [10]. This model employs spatiotemporal 3D convolution in all encoder and decoder modules to capture both the spatial and temporal components of the video sequence. Furthermore, in the recent MW-GAN+ network proposed by Jianyi Wang et al. [11], a 3D discriminator was also developed for improving the perceptual quality of compressed videos. The objective was to encourage temporal consistency using 3D-CNN-based architectures. Experiments indicate that the Generative Adversarial Network (GAN) framework, along with the 3D convolutional approach, can be adapted more efficiently and achieve stable time-consistent outcomes without further modifications for most datasets. Nonetheless, the inescapable computational cost tends to be higher, and adding more convolutional layers during training to achieve style migration with long videos could result in an increase in the number of parameters.

2.3. Introducing Inter-Frame Continuity in the Discriminator

One approach to addressing frame consistency issues in video style transfer is to introduce inter-frame continuity into the discriminator. The underlying principle is that we can modify the discriminator of the Generative Adversarial Network (GAN), so it not only assesses the authenticity of a single frame but also evaluates the continuity between successive frames. This encourages the generator to produce videos with consistent inter-frame transitions during the training process.

Chan et al. aimed to extend single-image generation, learning from a vast amount of personalized video data to synthesize temporally coherent and astonishingly realistic videos [12]. They altered the setup for single-image generation; the generator now predicts two consecutive frames, and the discriminator determines the continuity of these frames by comparing the "generated sequence" with the "original sequence", thereby addressing temporal consistency issues.

Similarly, Li et al. in their improvement of the image style pattern MGAN way, to prevent excessive consumption of computational resources, opted for a skip-frame generation approach [13]. Considering the continuity issues arising from skip-framing, they modified the discriminator to discern inter-frame continuity, thus ensuring temporal consistency of the frames. Meanwhile, Xu et al. proposed a time-consistent video semantic editing method, which also involves fine-tuning the generator to produce more continuous frames, significantly enhancing temporal consistency while preserving edited content [14]. Table 1 is the pros and cons of this approach.

Table 1. Introducing Inter-Frame Continuity in the Discriminator: Advantages and Disadvantages

Advantages	Disadvantages
Having the discriminator directly learn and discern the continuity between successive frames can guide the generator to better maintain inter-frame continuity.	The difficulty of training the model might increase.
During the training process, one can flexibly adjust the balance between style transfer quality and inter-frame continuity without the need for major modifications to the model structure or loss function.	There are strict requirements regarding the quantity of training data. In different training environments, it might be necessary to introduce new training strategies.
The complexity of the discriminator model is typically lower, which might lead to improved computational efficiency.	The dependency on the length of the frame sequence might increase.

2.4. Using A Time Loss Function

The temporal loss function is one of the methods used to address the frame consistency issue in video style transfer. Specifically, the purpose of the time loss function is to minimize visual discrepancies between consecutive frames, thereby ensuring visual continuity and reducing abrupt changes from frame to frame.

The time loss function is defined on account of the difference between style transfer results of the target frame and the original frame. The original frame can be either the preceding or following frame of the target frame. In this way, we can penalize visual discrepancies between consecutive frames, thus encouraging the model to generate results with greater frame consistency [15].

Ruder et al. proposed a method to transfer style from one image to an entire video sequence [15]. To avoid flickering in the video and discontinuity between frames, and to prevent the immense computational cost of processing each frame independently, they employed the temporal loss function. This function penalizes deviations in the trajectories generated between frames, pushing the model to produce videos where the style is consistent across consecutive frames. Although Cramer et al.'s study primarily focused on audio embedding, their paper touched upon the temporal loss function, which can be applied to video style transfer, ensuring synchronization between video motion and audio rhythm or content [16]. In this context, audio embedding provides contextual information about the audio, while the temporal loss function ensures that the video's temporal continuity aligns with audio information. Similarly, Yue et al. adopted a brand-new loss of motion consistency algorithm to intensify the continuity of motion between refactoring frames [17].

The general steps of these methods can be summarized as follows

- (1) For a pair of consecutive frames in a video, generate their respective style transfer results using a style transfer model.
- (2) Compute the difference between the results of these consecutive frames. This difference can be viewed as the temporal loss for this pair of consecutive frames.
- (3) Perform the aforementioned operations for all pairs of consecutive frames in the video. The sum of the temporal losses obtained is the total temporal loss for the video.
- (4) During the training process, the total temporal loss is used as a loss term to be minimized, optimized through backpropagation and optimizers.

Advantages and disadvantages of these methods are shown in table 2.

Table 2. Using the Temporal Loss Function: Advantages and Disadvantages

Advantages	Disadvantages
Reduce jitter and flickering.	Relatively dependent on the accuracy of optical flow computation.
Makes object movement smoother, without sudden changes or jumps.	Excessive continuity constraints might hinder the generation of innovative effects.
Ensures stability and continuity between consecutive frames in the video, enhancing video quality.	Computational complexity increases.
Enhance temporal consistency.	Might require more training data.

3. Data Comparison

3.1. Utilizing Recurrent Neural Networks (RNN)

Regarding the method that employs Recurrent Neural Networks (RNN), we analyze data from the paper by Xue et al [5], as an illustrative example. TO Flow represents a joint training module proposed by the authors that incorporates the RNN network. Table 3 compares the performance of different frame interpolation algorithms on two test sets. By examining the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) values on both test sets, the joint training module featuring the RNN network showcases the most exemplary performance across the three data metrics. Table 4 presents the quantitative outcomes of five frame interpolation algorithms on the Middlebury flow dataset, wherein, holistically, the TO Flow Mask yields the most superior results. In summary, it is evident that TO Flow, which incorporates the RNN network, outperforms other frame interpolation algorithms, affirming the distinct advantages of employing RNN. Furthermore, by comparing experimental data across several other papers, we discern that the RNN network model possesses a certain short-term memory attribute. The hidden layer information at each time is not only determined by the input layer at that time, but also affected by the hidden layer at the previous time. Consequently, employing the RNN network when addressing frame consistency issues results in superior fluidity and logical coherence.

Table 3. The performance of different frame interpolation algorithms on two test sets

Methods	Vimeo Interp		DVF Dataset	
	PSNR	SSIM	PSNR	SSIM
EpicFlow	32.0	0.962	33.7	0.964
DVF	33.2	0.962	34.1	0.963
AdaConv	32.3	0.957	—	—
SepConv	33.45	0.9674	34.69	0.9656
Fixed Flow	29.09	0.9229	31.61	0.9544
Fixed Flow + Mask	30.10	0.9322	32.23	0.9575
TOFlow	33.53	0.9668	34.54	0.9666
TOFlow + Mask	33.73	0.9682	34.58	0.9667

Table 4. Different frame interpolation algorithm in Middlebury data set on the test results

Methods	PMMST	DeepFlow	SepConv	TOFlow Mask
All	5.78	5.96	5.61	5.5
Discontinuous	9.55	9.78	8.74	8.5
Untextured	2.10	2.05	2.33	2.2

3.2. Employing 3D Convolution

Regarding the method of employing 3D convolution, given the varied applications in different papers, we take the research of Patraucean and colleagues as a representative example for analysis [8]. Table 5 compares the parameter count of various architectures and their average error per pixel.

Upon juxtaposition of the experimental data on average errors, it becomes evident that the model "ConLSTM-flow", into which the authors integrated 3D convolution operations, demonstrates a significantly reduced error rate in test datasets compared to the other three models. This suggests that the incorporation of 3D convolution plays a positive role overall. Collating experimental data from various other papers, we discern that models augmented with 3D convolution operations are adept at capturing both temporal and spatial feature information in videos, thereby enhancing the resultant stylized video in terms of quality and detail richness. Concurrently, 3D CNNs also exhibit superior performance levels in the realm of action recognition.

Table 5. Number of parameters and their average error for different architectures

Architecture	Parameters	Test error
Conv	109.070	0.0948
fcLSTM	33.623.650	0.0651
ConvLSTM	1,256303	0.0640
ConvLSTM-flow	1.035.070	0.0440

3.3. Introducing Inter-frame Continuity in the Discriminator

Regarding the method of introducing inter-frame continuity in the discriminator, we take the test results from Chan et al.'s paper as an exemplar for our analysis [12]. Within the table 6, FBF denotes videos obtained via direct frame-by-frame synthesis (control group), while FBF+TS signifies videos procured post the integration of frame continuity in the discriminator (experimental group). A comparative analysis of the experimental data for both reveals that volunteers involved in the study displayed a marked preference for videos resulting from the introduction of frame continuity. This underscores the positive impact of incorporating frame continuity in experiments, curtailing the emergence of artifacts, ensuring consistency between frames, and enhancing the generated video's overall quality. Experimental data from several other pertinent papers corroborate that introducing inter-frame continuity within the discriminator augments the smoothness of the generated videos, offering a more pleasurable viewing experience.

Table 6. Effect comparison of FBF, FBF+TS and FBF+TS+FG modules

Region	Metric	FBF	FBF+TS	FBF+TS+FG
Face	SSIM	0.784	0.811	0.816
	LPIPS	0.045	0.039	0.036
Body	SSIM	0.828	0.838	0.838
	LPIPS	0.057	0.051	0.050

Condition	vol1	vol2	vol3	vol4	vol5	Total
FBF	54%	70%	62%	54%	60%	59%
FBF+TS	60%	56%	50%	53%	53%	54%

3.4. Using a Temporal Loss Function

Regarding the method of utilizing a temporal loss function, we reference the results presented in Ruder et al.'s paper for an evaluation of its efficacy [15]. Table 7 provides a comparative analysis of short-term consistency benchmark results across various video styles for several methods. Both DeepFlow and EpicFlow incorporated the temporal loss function, while the other three methods employed random initialization without the inclusion of this loss function. The outcomes reveal that in most scenarios, the use of the temporal loss function (also referred to as temporal penalty) enhanced the results. However, an anomaly was observed in the "ambush" scene. This deviation can be attributed to the intense movement characteristic of an ambush scenario, causing optical flow computations to err, thus undermining frame consistency regardless of whether the temporal loss function was applied. This highlights the method's limitation, particularly its heavy reliance on the

accuracy of optical flow computations. The method's efficacy diminishes in scenes characterized by intense motion. Moreover, the necessity to continuously compute inter-frame differences due to the loss function leads to an uptick in computational resource requirements.

Table 7. Four methods for short-term consistency of benchmark results in different video styles

	Sence1	Sence2	Sence3	Sence4	Sence5
Deep Flow	0.0006	0.006	0.012	0.0008	0.0035
Epic Flow	0.0007	0.007	0.014	0.0008	0.003
Init prev	0.01	0.02	0.028	0.004	0.014
Init random	0.02	0.03	0.037	0.02	0.02

4. Conclusion

This paper has meticulously analyzed mainstream solutions for addressing frame consistency issues during the video style transformation process, focusing specifically on methods rooted in Generative Adversarial Networks (GAN). Drawing upon a wealth of literature, we have synthesized and elucidated four predominant GAN-based strategies to resolve the frame consistency challenges. The characteristics of these methods were explored, and through the gathered experimental data, the advantages and drawbacks of each approach were discerned. This study serves to augment the existing body of knowledge within the research domain of video style transfer.

Although there are already some successful practical applications of video style transfer in everyday life, the field still confronts a host of technical challenges. The inherent complex structure of video makes its generation a daunting task that requires a lot of time and effort to train the model for subsequent output. At the same time, the computational requirements for training these models are very high. The size and quality of the data sets used to train these models affect their complexity.

Given these outstanding challenges, video style transfer remains in its developmental phase. While the models discussed earlier have achieved commendable outcomes, the application of these experimental results in real-world scenarios is still fraught with impediments. With this in mind, we project several directions for future research in this area:

1. Generating Higher-Resolution Videos: While the aforementioned models have successfully produced results, there remains room for quality enhancement. A potential reason could be the subpar quality of datasets (comprising low-resolution samples) used for model training.

2. Optimization of data sets: most contemporary works using synthetic data sets or with limited movement of the real data sets. Therefore, there is an urgent need to apply these models to more realistic datasets.

3. Reduce the complexity of the network: in order to minimize the time needed for processing video, usually using RNN model and deconvolution, the two methods are introduced too many parameters, thus improve the complexity of the network. Therefore, compared with the existing models, fewer parameters are introduced methods to solve the problem of generating time more effectively is imperative.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Thimonier, H. et al. Learning Long Term Style Preserving Blind Video Temporal Consistency. 2021 IEEE International Conference on Multimedia and Expo, 2021, 1 - 10.
- [2] Lai, W.-S., Huang, J.-B., Wang, O., Shechtman, E., Ersin Yumer, & Yang, M.-H. Learning Blind Video Temporal Consistency. European Conference on Computer Vision, 2018: 179 – 195.

- [3] Patil, P. W., Dudhane, A., & Murala, S. End-to-End Recurrent GAN for Traffic and Surveillance Applications. *IEEE Transactions on Vehicular Technology*, 2020, 69 (12): 14550 - 14562.
- [4] Xiang, L., Kong, G., Duan, X., Long, H., & Wu, Y. CRVC-GAN: combining cross-scale fusion and recursion for video colourization GAN. *Journal of Electronic Imaging*, 2022, 31 (06).
- [5] Xue, Tianfan, et al. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision*, 2019, 127 (8): 1106 – 1125.
- [6] Brouwer, E.D., Simm, J., Arany, A., & Moreau, Y. GRU-ODE-Bayes: Continuous modelling of sporadically-observed time series. *neural Information Processing Systems*, 2019, 32: 7377 - 7388.
- [7] Chen, Xin, et al. Singing Voice Conversion with Non-parallel Data. *2019 IEEE Conference on Multimedia Information Processing and Retrieval*, 2019: 292 - 296.
- [8] Patraucean, V., Handa, A., & Cipolla, R. Spatio-temporal video autoencoder with differentiable memory. *arXiv (Cornell University)*, 2015.
- [9] Yan, Mengjia, et al. 3D Convolutional GAN for Detecting Temporal Irregularities in Videos. *2018 24th International Conference on Pattern Recognition*, 2018.
- [10] Aigner, S., and M. Körner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3D convolutions in progressively growing gans. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2019, XLII-2/W16: 3 – 11.
- [11] Wang, Jianyi, et al. MW-Gan+ for perceptual quality enhancement on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32 (7): 4224 – 4237.
- [12] Chan, Caroline, et al. Enhancing Space-time Video Super-resolution *2019 IEEE/CVF International Conference on Computer Vision*, 2019: 5932 - 5941.
- [13] Li, Wenbo, et al. Evolvment Constrained Adversarial Learning for Video Style Transfer. *Computer Vision, Asian Conference on Computer Vision*, 2018, 2019: 232 – 248.
- [14] Xu, Yiran, et al. Temporally consistent semantic video editing. *Lecture Notes in Computer Science*, 2022: 357 – 374.
- [15] Ruder, Manuel, et al. Artistic style transfer for videos. *Lecture Notes in Computer Science*, 2016: 26 – 36.
- [16] Cramer, Aurora Linh, et al. Look, listen, and learn more: Design choices for Deep Audio Embeddings. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: 3852 – 3856.
- [17] Z. Yue and M. Shi, Enhancing Space-time Video Super-resolution via Spatial-temporal Feature Interaction, *arXiv (Cornell University)*, 2022.