Video reconstruction of generative adversarial network based on global perception

Chen Shende¹, Wang Xiaoman¹ and Jing Wenbo²
¹College of Electronic Information and Engineering, Changchun University of Science and Technology, Changchun, 130022
²College of Optoelectronic Engineering, Changchun University of Science and Technology, Changchun, 130022
E-mail: huangbingkun@mails.cust.edu.cn

Abstract. Aiming at the problem that the current video reconstruction can not take into account the high perception quality and high temporal correlation, a video reconstruction of generative adversarial network based on global perception is proposed. The generator adopts a recurrent and iterative network architecture. In order to make up for the lack of extraction of global information by the frame alignment network, the alignment network combines the global information extracted by the global information perception to perform frame alignment. Through the static temporal loss and the temporal statistics loss, combined with the relative discriminator of sequence frames input, the temporal correlation of the generated images sequence is improved. Experimental comparisons were performed on the Vid4 and REDS test sets, and the best image perception quality index (LPIPS/NIQE) were obtained, as well as better temporal correlation (tLPIPS), reaching 0.192/3.417/0.328 and 0.138/3.223/0.217. Experimental results show that the proposed method can effectively improve the perceptual quality of reconstructed video frames and has good temporal correlation.

1. Introduction
With the wide application of remote sensing satellite, face recognition and medical imaging technology, video super-resolution reconstruction has been applied to various fields of image processing and machine vision tasks. In recent years, video super-resolution reconstruction based on deep learning has become the mainstream method, and the images reconstructed by using generative adversarial network have more natural and realistic texture details. Among them, SRGAN[1] pioneered the use of generative adversarial networks to reconstruct high-resolution images, and proposed a perceptual loss function, which significantly improved the perceptual quality of the generated images. EnhanceNet[2] adopts a fully convolutional network structure and improves the loss function, which makes the texture of the generated images more realistic and effectively reduces unrealistic artifacts. The VSRResFeatGAN[3] generative adversarial network optimizes the generator part of SRGAN and proposes a new discriminator architecture to obtain higher perceptual quality video images in video reconstruction. In order to solve the problem of instability in the training of the generative adversarial network, WGAN[4] proposes a generator loss and a discriminator loss that minimizes the Wasserstein distance between the generated distribution and the true distribution, and improves the stability of network training through gradient truncation. ESRGAN[5] introduces the residuals in the residual dense block as the basic unit of the reconstruction network, removes batch normalization, and improves the generation of resistance loss and perceptual loss, and obtains better image perception quality on more realistic and natural textures.
Researchers such as Eduardo[6] proposed static temporal loss and temporal statistics loss to improve the temporal correlation of reconstructed video sequence frames. However, although the current deep learning network has significantly improved the quality of reconstructed video images, most of them are trained by pixel error driven method. Although a higher peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index and a better time correlation (tLPIPS) index can be obtained, but at the same time, it will also bring about the blur artifacts that the edge texture is too smooth, which will affect the image perception of human vision. While the network trained in the perceptual drive mode can obtain higher perceptual quality images, but the current video images reconstructed in the perceptual drive mode are relatively poor in temporal correlation, and it is easy to cause inter-frame flicker in the video. It has also become a problem to be improved in the reconstruction of video super-resolution.

To solve this problem, a generative adversarial network video reconstruction based on global perception (VRGPGAN) is proposed. In this paper, the optical flow estimation alignment network is improved and combined with the global features extracted by global information perception to align the two adjacent frames before and after, and the alignment features containing global and local information are obtained. The relative discriminator that uses the sequence frames input determines the true probability of generating the sequence frames from the temporal and spatial information, so that the video frames generated by the generator has a more real temporal correlation. In order to better eliminate the flicker between the reconstructed video frames, the static temporal loss and the temporal statistics loss are introduced to constrain the training generator to enhance the temporal correlation of the reconstructed video frames.

2. Method

The generator architecture of recurrent iteration is adopted, and the reconstruction output of the previous time period of the iteration is used as the reconstruction input of the next time period. The video sequence frames is used as the input of the discriminator, and a small video segment of the sequence frames is discriminated each time. Taking the discrimination of 3 sequence frames each time as an example, the high-resolution current frame $I^R_t$ reconstructed each time, the high-resolution frames reconstructed in the previous two time periods, and the corresponding real high-resolution frames are input as the discriminator. The overall architecture of VRGPGAN is shown in Figure 1.

![Figure 1. Overall architecture of VRGPGAN.](image)

2.1. Generator architecture

As shown in Figure 2, the generator architecture is composed of global information perception, alignment network, and reconstruction network, iterating the high-resolution frame $I^R_{t-1}$ reconstructed in the previous time period, combining the current frame $I^L_t$ and the low-resolution frame $I^L_{t-1}$ of the
previous frame as the input of the generator. The alignment network uses the optical flow estimation network to estimate the motion information of the current frame and the previous frame, and combines the global information extracted by global information perception to perform frame alignment. The alignment feature $I_{t}^{'}$, $M_{t-1}^{context}$, the global perception feature $I_{t}^{context}$ of the current frame and the estimated optical flow motion information $F_{t}$ of the current frame are concatenated and fed into the reconstruction network. The reconstruction network in the literature[7] is adopted, and the input current frame is up-sampled bilinearly 4 times and then jumped to the output of the reconstruction network to accelerate the convergence speed of network training.

2.1.1. Global information perception network. The global information perception network is embedded with the GIA module of reference[8], as shown in Figure 3, where k is the size of the convolution kernel, n is the number of channels, and s is the step size. In order to achieve the effect of global information perception, the input low-resolution image is convolved to obtain a feature map $X$ of size $H \times W \times C$; The global information is extracted through the down-sampling function $f_{1}(X)$ to generate a feature map of $1 \times 1 \times C$; then the up-sampling function $f_{2}(X_{i})$ is used to restore the size of the down-sampled input feature, after Conv$(1 \times 1)$ Convolution reduces the number of channels and combines with $X$, and finally through Conv $(1 \times 1)$ convolution processing to obtain a feature map containing local information and global information. The global information $I_{t}^{context}$ of the current frame is obtained from the current frame $I_{t}^{LR}$ through the global information perception network. For the global information of the previous frame, $M_{t-1}^{LR}$ is obtained by feature extraction of the reconstructed image of the previous stage, and then $M_{t-1}^{LR}$ and the previous frame $I_{t-1}^{LR}$ are input into the global information perception network to obtain the global information $M_{t}^{context}$ of the previous frame.

2.1.2. Frame alignment network. Assuming that the optical flow is a dense optical flow, each pixel is allowed to be transformed to a new position, so that the resulting pixel arrangement can be aligned to a regular grid using bilinear interpolation $\text{Warp}\{\}$. According to the literature[9], the frame alignment of optical flow compensation can be described by the following formula:
\[ I'_t = \text{Warp}\{I_{t}^{LR}(F_t)\} \tag{1} \]

Where \( F_t \) is the optical flow estimation feature of the current frame. The frame alignment network architecture is shown in Figure 4. The optical flow motion estimation and compensation in VESPCN[9] is improved in this paper. The optical flow estimation network is shown in Table 1, which is divided into coarse optical flow estimation and fine optical flow estimation. The frame alignment network estimates the optical flow feature \( F_t \) of the current frame \((t \rightarrow t)\) and the optical flow feature \( F_{t-1}^{R} \) of the previous frame \((t \rightarrow (t-1))\) respectively. The current frame \( I_{t}^{LR} \) and the previous frame \( I_{t-1}^{LR} \) are input to the coarse optical flow estimation network to obtain the coarse estimated optical flow feature \( F_{t-1}^{c} \), and then perform image alignment between \( F_{t-1}^{c} \) and the current frame \( I_{t}^{LR} \) to obtain the coarse alignment feature \( I_t^{c} \). Then the original input image, coarse alignment feature \( I_t^{c} \) and coarse estimated optical flow feature \( F_{t-1}^{c} \) are used as the input of the fine optical flow estimation network to obtain a more refined optical flow feature \( F_t^{f} \) of the current frame. The final current frame alignment feature \( I'_t \) is obtained by the motion compensation of the current frame optical flow feature \( F_t^{f} = F_{t-1}^{c} + F_t^{f} \) to the current frames \( I_{t}^{LR} \) and \( I_{t-1}^{context} \). Similarly, the alignment feature \( M_{t-1}^{context} \) of the previous frame is obtained by the motion compensation of the optical flow feature \( F_{t-1}^{R} = F_{t-1}^{c} + F_{t-1}^{f} \) of the previous frame to the previous frames \( I_{t-1}^{LR} \) and \( M_{t-1}^{context} \), which can be described by the following formulas:

\[ I'_t = \text{Warp}\{I_{t}^{LR}(F_t) + I_{t}^{context}(F_t)\} \tag{2} \]
\[ M_{t-1}^{context} = \text{Warp}\{I_{t-1}^{LR}(F_{t-1}) + M_{t-1}^{context}(F_{t-1})\} \tag{3} \]

### Figure 4. Frame alignment network architecture.

### Table 1. Optical flow estimation network.

| Layer | Coarse flow          | Fine flow          |
|-------|----------------------|--------------------|
| 1     | Conv k5n24s2/ReLU    | Conv k5n24s2/ReLU  |
| 2     | Conv k3n24s1/ReLU    | Conv k3n24s1/ReLU  |
| 3     | Conv k5n24s2/ReLU    | Conv k3n24s1/ReLU  |
| 4     | Conv k3n24s1/ReLU    | Conv k3n24s1/ReLU  |
| 5     | Conv k3n32s1/ReLU    | Conv k3n8s1/tanh   |
| 6     | Sub-pixel upscale×4  | Sub-pixel upscale×2|
2.2. Sequence frames discriminator

Because the discriminator of single frame input cannot use the temporal information between sequence frames, the discriminator of single frame input in SRGAN[1] is used as the network architecture, and it is improved to the discriminator of sequence frames input, as shown in Figure 5. Two sets of sequence frames are used as the input of the discriminator, which are the high-resolution sequence frames \( I_{t-n}, \ldots, I_t \) reconstructed by the generator and the corresponding real high-resolution sequence frames \( I_{t-n}, \ldots, I_t \). Using sequence frames input helps the discriminator to identify the true and false of the generated sequence frames closer to the real sequence frames.

Figure 5. Sequence frames discriminator.

2.3. Losses

Inspired by the single frame relative discriminator of ESRGAN[5], the sequence frames relative discriminator \( D(\cdot) \) is used in this paper to try to predict the probability that the real sequence frames \( I = \{ I_{t-n}, \ldots, I_t \} \) is more realistic than the generated sequence frames \( I^{HR} = \{ I_{t-n}^{HR}, \ldots, I_t^{HR} \} \).

\[
D(I, I^{HR}) = \text{sigmoid}(C(I) - C(I^{HR})) \rightarrow 1
\]

\[
D(I^{HR}, I) = \text{sigmoid}(C(I^{HR}) - C(I)) \rightarrow 0
\]

Where \( C(\cdot) \) is the output of the discriminator without non-linear transformation. The discriminator loss \( L_D \) and the generator confrontation loss \( L_G \) are respectively defined as:

\[
L_D = -\log(D(I, I^{HR})) - \log(1 - D(I^{HR}, I))
\]

\[
L_G = -\log(1 - D(I, I^{HR})) - \log(D(I^{HR}, I))
\]

In order to accurately obtain the edge and contour information, the \( L_1 \) norm of the MSE loss is used to minimize the pixel distance between the generated frame and the real frame, and \( L_{MSE} \) is defined as:

\[
L_{MSE} = \left\| I_t^{HR} - I_t \right\|
\]

In order to match the texture of the generated image to the real image texture during training, the full-text information loss \( L_{CI} \) with global information in the literature[10] is used:

\[
CI(I_t, I_t^{HR}) = \frac{1}{N} \sum_{j} \max_{i} CI_{ij}
\]

\[
L_{CI}(I_t, I_t^{HR}, I) = -\log(CI(\phi(I_t), \phi(I_t^{HR})))
\]
Where $CI_y$ is the calculation of the similarity between the $x_i \in I_i$ feature and the $y_j \in I_{i,IR}$ feature, and $\phi(\cdot)$ is the feature extracted by the $l$ layer of the VGG-19 network.

In order to enhance the temporal correlation between the reconstructed video frames, static temporal loss $L_{t_j}$ and temporal statistics loss $L_{t_i}$ in literature[6] are introduced:

$$m_i = \exp(-\alpha \| l_i - l_{i-1} \|^2)$$  \hspace{1cm} (11)

$$L_{t_j} = m_i \| I_{i,IR} - I_{i+1,IR} \|^2$$  \hspace{1cm} (12)

$$L_{t_i} = \| \sigma^2 - \sigma_{i,IR}^2 \|^2$$  \hspace{1cm} (13)

Where $m_i$ is the mask calculated using the real image, $\alpha$ is 100 by default, $\phi$ is the mask operation, $\sigma^2$ is the variance of each position $I_i(x,y)$ of the real image, and $\sigma_{i,IR}^2$ is the variance of each position $I_{i,IR}(x,y)$ of the generated image.

3. Experiment and result analysis

3.1. Data set and parameter settings

Network training uses REDS[11] training set and Vimeo-90K training set, and Vid4 test set and REDS test set are used for testing. REDS is a newly established video data set with realistic and diverse scenes in the NTIRE19 competition. Vimeo-90K contains high-resolution videos of various indoor and outdoor scenes. Due to the lack of corresponding low-resolution videos, linear down-sampling and noise addition methods are used to generate corresponding low-resolution video frames.

Use the combined loss $L$ to train the generator:

$$L = \lambda_1 L_g + \lambda_2 L_{mse} + \lambda_3 L_{tc} + \lambda_4 L_{t_j} + \lambda_5 L_{t_i}$$  \hspace{1cm} (14)

The weight of each loss function is $\lambda_1=0.005$, $\lambda_2=0.01$, $\lambda_3=1$, $\lambda_4=0.1$, $\lambda_5=0.1$. For the weighting ratio, each loss function is trained independently, and the magnitude of each loss function is observed when it is about to converge, and the reference weight is determined. Set the training mini-batch size to 24, crop the input low-resolution video frame size to 64×64, set the learning rate to $1 \times 10^{-4}$, use the Xaier method to initialize the weights, and adjust the learning rate through annealing.

3.2. Analysis of results

In this paper, the peak signal-to-noise ratio (PSNR), structural similarity (SSIM), image perception similarity (LPIPS)[12], natural image quality evaluation (NIQE)[13] and the image perception distance between generated continuous frames and real continuous frames (tLPIPS)[6] Five indicators respectively evaluate the quality of reconstructed video images. The lower the LPIPS index and the NIQE index, the better the perceived quality of the generated images. The lower the tLPIPS index, the better the temporal correlation between the generated continuous frames and the real continuous frames in image perception similarity.

The reconstruction images quality evaluation of VRGPGAN and other models on the test set Vid4 and REDS is shown in Table 2 and Table 3. The bolded indicates the best. It can be seen that the VRGPGAN method in this paper has obtained the best LPIPS and NIQE indexes, has the best image perception quality, and also obtained better PSNR, SSIM and tLPIPS indexes. Compared with DUF[14], because DUF is driven by pixel error, and the frame alignment network uses implicit motion compensation, it avoids the requirement for accurate optical flow motion estimation. In addition, DUF enhances the texture information of the reconstructed images by generating up-sampling kernel and residual network, so it has higher PSNR, SSIM and tLPIPS indexes. However, this pixel error driving method that only uses MSE as a loss function can obtain a higher PSNR index, but it will also bring fuzzy texture artifacts and reduce the perceptual quality of the generated images[15]. It can be seen from
the table that DUF driven by pixel error is worse than ESRGAN, EnhanceNet and VRGPGAN driven by perception in both LPIPS and NIQE indices.

| Method / Index | PSNR  | SSIM  | LPIPS | NIQE  | tLPIPS |
|---------------|-------|-------|-------|-------|--------|
| ESRGAN        | 21.892| 0.687 | 0.241 | 3.507 | 0.498  |
| EnhanceNet    | 20.886| 0.683 | 0.253 | 6.470 | 0.504  |
| DUF           | **26.811**| **0.815**| 0.268 | 6.923 | **0.149**|
| VRGPGAN       | 25.899| 0.784 | **0.192**| **3.417**| 0.328  |

Table 3. Quality evaluation of reconstructed images in the REDS test set.

| Method / Index | PSNR  | SSIM  | LPIPS | NIQE  | tLPIPS |
|---------------|-------|-------|-------|-------|--------|
| ESRGAN        | 26.03 | 0.804 | 0.196 | 3.574 | 0.296  |
| EnhanceNet    | 25.831| 0.815 | 0.213 | 4.970 | 0.383  |
| DUF           | **28.631**| **0.875**| 0.237 | 5.622 | **0.174**|
| VRGPGAN       | 27.502| 0.866 | **0.138**| **3.223**| 0.217  |

Figure 6 lists the reconstructed images effects of VRGPGAN and different models. It can be seen that VRGPGAN reconstructs more realistic and natural detailed textures and edge contours. The details of building edges, text textures and wheel outlines in the reconstructed video images are clear and identifiable, with better image perception quality.
Figure 7 shows the effects of sequence frames changes reconstructed by VRGPGAN and different reconstruction models, reflecting the comparison in temporal correlation. HR is real high-resolution sequence frames. It can be seen that the sequence frames reconstructed by VRGPGAN and the real high-resolution sequence frames have a good change consistency. The DUF uses a pixel error-driven method to train the network. Although it can obtain a good change consistency, the reconstructed texture produces blurry artifacts. Although both ESRGAN and EnhanceNet adopt a perception-driven approach to improve images perception quality, the generated sequence frames are not consistent with the real sequence frames.

![Figure 7. Reconstruction sequence frames change effects.](image)

### 3.3. Analysis of ablation experiments

In order to verify the effectiveness of global information perception and the effectiveness of the method in this paper on temporal correlation constraint. The VRGPGAN network with global information perception removed, and the VRGPGAN network with static temporal loss and temporal statistic loss removed were tested for ablation experiments. The reconstruction images quality evaluation of related experiments is shown in Table 4. RL_{C1} means that global information perception is removed, and RL_{T} means that static temporal loss and temporal statistical loss are removed.

It can be seen from the table that the various indicators of RL_{C1} show obvious degradation, indicating that the global information extracted by global information perception can effectively enhance the frame alignment accuracy and improve the perceived quality of the reconstructed images. The various indicators of the RL_{T} are also degraded as a whole, which shows that the static temporal loss and the temporal statistics loss have an effective constraint on the temporal correlation of the reconstructed video frames. At the same time, the RL_{T} index is compared with Table 2, and RL_{T} has obtained better tLPIPS index than ESRGAN and EnhanceNet. It shows that the generator architecture using recursive iteration and the relative discriminator of the sequence frames input can also effectively improve the temporal correlation of the reconstructed sequence frames.

| Method / Index | PSNR | SSIM | LPIPS | NIQE | tLPIPS |
|----------------|------|------|-------|------|--------|
| RL_{C1}        | 24.133 | 0.725 | 0.218 | 3.485 | 0.336  |
| RL_{T}         | 25.352 | 0.738 | 0.247 | 3.744 | 0.379  |
| VRGPGAN        | 25.899 | 0.784 | 0.192 | 3.417 | 0.328  |

### 4. Conclusions

This paper studies how to improve the accuracy of frame alignment and the temporal correlation of reconstructed video frames, and a video reconstruction method of generative adversarial network based on global perception is proposed. The studies have achieved the following three achievements: First, in terms of frame alignment, a lightweight and effective global information perception is introduced, which makes up for the lack of global information extraction in the previous frame alignment network. Secondly, in terms of temporal correlation, the sequence frames relative discriminator is proposed, which solves the problem that the single frame relative discriminator can not use the temporal information of video frames. In addition, this paper constructs a recursive iterative generator network, and introduces static temporal loss and temporal statistical loss to constrain the network training, which
further strengthens the temporal correlation of reconstructed video frames. Experimental results and ablation experiments show that this method can not only effectively improve the image perception quality of reconstructed video frames, but also obtain good temporal correlation.

References
[1] Christian L, Theis L, Ferenc H, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z and Shi W 2017 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI, USA 105-14
[2] Sajjadi M S M, Bernhard S and Michael H 2017 Proceedings of the IEEE International Conference on Computer Vision (Venice, Italy: IEEE) 4501-10
[3] Alice L, Santiago L, Rafael M and K K A 2019 IEEE T IMAGE PROCESS 28 3312-27
[4] Martin A, Soumith C and Léon B 2017 arXiv:1701.07875 2 156-63
[5] Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y and Loy C C 2019 ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks (Cham: Springer International Publishing)
[6] Eduardo P, M S M S, Michael H and Bernhard S 2018 Computer Vision and Pattern Recognition
[7] Niklaus S and Liu F 2018 (Salt Lake City, UT, USA: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition) 1701-10
[8] Meng Z, Xu R and Ho C M 2020 Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12537 LNCS 327-42
[9] Caballero J, Christian L, Aitken A, Acosta A, Totz J, Wang Z and Shi W 2017 Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 20172017-January 2848-57
[10] Mechrez R, Talmi I and Zelnik-Manor L 2018 European Conference on Computer Vision 800-15
[11] Wang X T, Chan K C K, Yu K, Chao D and Chen L C 2019 (Long Beach, CA, USA 1954-63
[12] Zhang R, Isola P, A E A, Shechtman E and Wang O 2018 arXiv:1801.03924: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)) 586-95
[13] Mittal A, Soundararajan R and Bovik A C 2013 IEEE SIGNAL PROC LET 20 209-12
[14] Jo Y, Oh S W, Kang J and Kim S J 2018 (Salt Lake City, UT, USA: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition) 3224-32
[15] Ma C, Rao Y, Cheng Y, Chen C, Lu J and Zhou J 2020 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Seattle, WA, USA: IEEE) 7766-75