FAPN: Face Alignment Propagation Network for Face Video Super-Resolution

Sige Bian[1], He Li[1], Feng Yu[2], Jiyuan Liu[1], Song Changjun[1], and Yongming Tang[1](✉)

[1] Joint International Research Laboratory of Information Display and Visualization, Southeast University, Nanjing 210096, China
{220211638,erie,scj,tym}@seu.edu.cn, he.li@ieee.org

[2] School of Computing, National University of Singapore, Singapore, Singapore
yufeng@nus.edu.sg

Abstract. Face video super-resolution (FVSR) aims to use continuous low resolution (LR) video frames to reconstruct face and recover facial details under the premise of ensuring authenticity. The existing video super-resolution (VSR) technology usually uses inter-frame information to achieve better super-resolution (SR) performance. However, due to the complex temporal dependence between frames, as the number of input frames increases, the information cannot be fully utilized, and even wrong information is introduced, resulting in poor performance. In this work, we propose an alignment propagation network for accumulating facial prior information (FAPN). We design a neighborhood information coupling (NIC) module based on optical flow estimation and alignment, where the current frame, the adjacent frames and the SR results of the previous frame are locally fused. The coupled frames are sent to a unidirectional propagation (UP) structure for propagation. Meanwhile, in the UP structure, the facial prior information is filtered and accumulated in the face super-resolution cell (FSRC), and the high-dimensional hidden state is introduced to propagate effective temporal information between frames along the unidirectional structure. Extensive evaluations and comparisons validate the strengths of our approach, FAPN can accumulate more facial details while ensuring the authenticity of the face. And the experimental results demonstrated that the proposed framework achieves better performance on PSNR (up to 0.31 dB), SSIM (up to 0.15 dB) and face recognition accuracy (up to 1.99%) compared with state-of-the-art methods.

Keywords: Face video super-resolution · Alignment propagation network · Face recognition accuracy

1 Introduction

With the development of artificial intelligence technology, face video super-resolution (FVSR) has been widely used in intelligent transportation, personal

This work is supported by National Natural Science Foundation of China (grant number 62275046).

Y. Zheng et al. (Eds.): ACCV 2022, LNCS 13848, pp. 3–18, 2023.
https://doi.org/10.1007/978-3-031-27066-6_1

identification, public security and other fields [1]. In the field of video surveillance, due to the hardware limitations of video capture equipment, it is difficult to obtain clear frames of target face far away from equipment [2]. FVSR aims to effectively enhance facial details in video [3], improve the accuracy of face recognition, and provide greater utilization of raw data (Fig. 1).

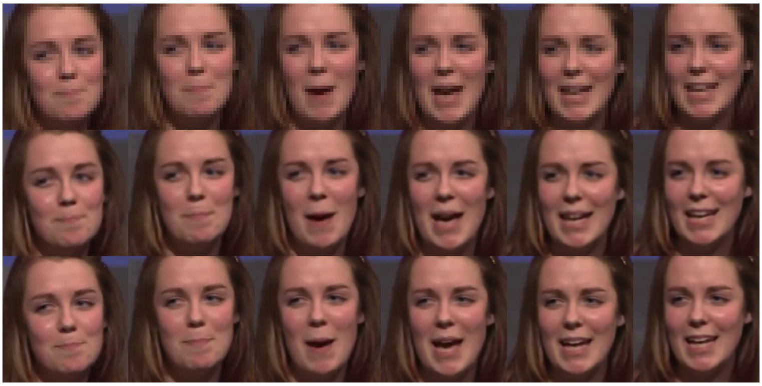

Fig. 1. Visual results of our FAPN on scale factor 4. Six consecutive LR video frames (1st row), HR frames (2st row, generated by our method) and groundtruth (GT) frames (3st row) are shown.

Compared with single image super-resolution (SR), video super-resolution (VSR) can achieve better performance by utilizing inter-frame information. At present, there are two main categories to utilize inter-frame information, alignment-based and non-alignment-based. RBPN [4] integrates the spatial and temporal context of consecutive video frames using a recurrent encoder and decoder module in the multi-projection stage. EDVR [5] uses deformable convolution to complete frame alignment at the feature level in a coarse-to-fine manner. The alignment-based approach is efficient, but it only uses information from adjacent frames and cannot effectively use input information far from the current frame. To effectively utilize more inter-frame information, non-alignment-based approaches [6–8] have also been proposed. Although these recurrent structures can accumulate more information from frames, it will inevitably introduce interference or even erroneous information useless for the current frame SR.

In the aspect of face super-resolution (FSR), rational use of strong constraints of human face will bring abundant prior information to the SR process. In addition, if a reasonable information accumulation mechanism is used to accumulate correct and useful information for the face region in the video, the details of the face region can be effectively enhanced under the premise of ensuring the authenticity of the generated face.

In this work, we propose a face video super-resolution network (FAPN), which combines the advantages of alignment-based and non-alignment-based methods. The current frame, the adjacent frames and the SR results of the previous frame are locally fused, and then sent to a unidirectional propagation (UP) structure for propagation, which is based on optical flow estimation and alignment. In addition, in the propagation structure, we filter and accumulate face information, and introduce hidden state to propagate the effective information forward with UP structure. FAPN outperforms existing state of the arts in PSNR (up to 0.31 dB), SSIM (up to 0.15 dB) and face recognition accuracy (up to 1.99%).

Our main contributions can be summarized as follows: 1) We redesigned the inter-frame alignment structure and the propagation structure so that the coupled inter-frame information could be propagated between frames after neighborhood information coupling (NIC) module, thus improving the accuracy of the model. 2) We propose an effective facial prior information filtering mechanism, retain correct and effective information, realize the accumulation of face information to enrich face details. 3) Combined with the prior information of human face, we introduce the pixel loss of facial features and the loss of high-level feature vectors of face to constrain the network training process.

2 Related Work

2.1 Video Super-Resolution

Alignment-based VSR methods mainly include motion compensation and deformable convolution. For motion compensation, VESPCN [9] consists of an alignment network and a fusion spatiotemporal sub-pixel network, which can effectively utilize temporal redundancy, but with low accuracy of generated images. EDVR [5] achieves frame alignment at the feature level in a coarse-to-fine fashion using deformable convolutions. Alignment-based methods usually utilize image information of adjacent frames but cannot effectively utilize input information far from the current frame.

As for non-alignment-based approaches, DUF [6] generates dynamic upsampling filters and residual images based on the local spatiotemporal neighborhood of each pixel to avoid explicit motion compensation. This method is computationally intensive and will introduce memory burden. RLSP [8] introduces a recurrent structure, where information propagates through hidden states. BasicVSR [10] combines feature-level alignment with the bidirectional propagation mechanism, which has a good performance improvement. However, the bidirectional propagation mechanism needs to acquire all the video sequences before VSR, which is not easy to be applied in real-time VSR.

As for face VSR, it aims to improve the resolution of facial regions. Xin [11] proposes a Motion Adaptive Feedback Unit (MAFC) that filters out unimportant motions such as background motions, preserves facial normal rigid motions and non-rigid motions of facial expressions, and feeds them back to the network. Yu [3] optimized the network parameters through three search strategies of TPE, random search and SMAC, and proposed a lightweight FVSR network HO-FVSR. Different from previous works, we propose an end-to-end FVSR framework (FAPN) to accumulate correct facial information. It is demonstrated that our face information accumulation mechanism facilitates our framework to achieve the state-of-the-art performance.

2.2 Facial Information Extraction and Recognition

Different from general VSR, in terms of FVSR, we mainly focus on the authenticity of the generated face, with the purpose of improving the accuracy of face recognition. In the field of face recognition, high-dimensional feature vectors are usually used to represent facial information. Euclidean distance is calculated for the generated feature vectors. When Euclidean distance becomes smaller, the face similarity increases. Representative methods in this field include Openface [12], Face_recognition [13], Insightface [14] and other networks. The accuracy of Face_recognition and Insightface is 99.38% and 99.74% respectively, which can be used for feature extraction and result evaluation.

3 Network Architecture

3.1 Framework of FAPN

VSR aims to map a LR video sequence $I^{LR} \in \mathbb{R}^{H\times W\times C}$ to a HR video sequence $I^{HR} \in \mathbb{R}^{kH\times kW\times C}$, where k is the upsampling factor, H and W are the height and width, and C is the number of channels. We propose a Face Alignment Propagation Network (FAPN) to filter and accumulate facial information. Our network structure is shown in Fig. 2.

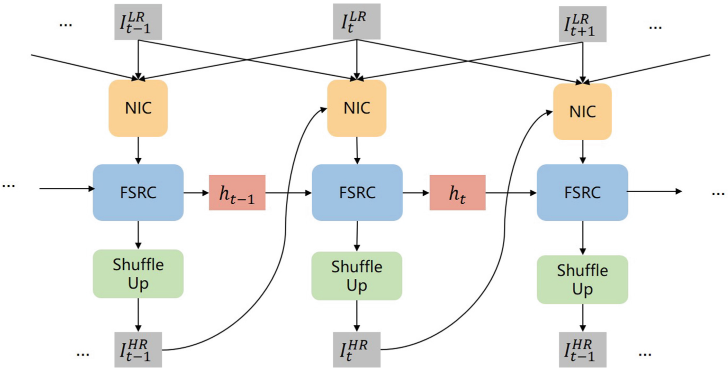

Fig. 2. Overview of the proposed framework FAPN.

Our framework takes consecutive LR frames $I_t^{LR} \in \mathbb{R}^{H\times W\times C}$ $(t = 0, 1, 2\ldots)$ as input to the network. We design a neighborhood information coupling (NIC) module to couple the information of adjacent frames in the first stage. In addition, in order to utilize previous SR results, feedback is introduced, thus I_{t-1}^{HR} is also sent to the NIC module at time t. In the second stage, the coupled information is sent to a unidirectional propagation (UP) structure. By introducing the hidden state $h_t(t = 0, 1, 2...)$, the UP structure can propagate information from the first frame to the current frame to supplement information. FSRC stands for facial information SR module, which is used to extract facial prior information in the network and generate face video streams with rich details and high authenticity. At time t, the output of FSRC are hidden state h_t and the filtered facial features I_t^{FSR}. I_t^{FSR} is shuffled up to get the final output.

3.2 Neighborhood Information Coupling

Our NIC module aims to add details to the current frame. It is first coupled with the information of adjacent frames and previous SR results, then the coupled information is sent to the propagation structure. The NIC module is mainly composed of optical flow estimation (OFE) module, alignment module and shuffling module (see Fig. 3).

The role of OFE is to predict HR optical flow from LR frame, and the predicted HR optical flow can be used to align the frames in the neighborhood with the current frame, which helps to reconstruct more accurate temporal information. In NIC, we adopt OFRnet [15] for optical flow information estimation in OFE module due to its simplicity and effectiveness:

$$OF_{t-1}^{HR} = OFE(I_{t-1}^{LR},\ I_t^{LR}) \tag{1}$$

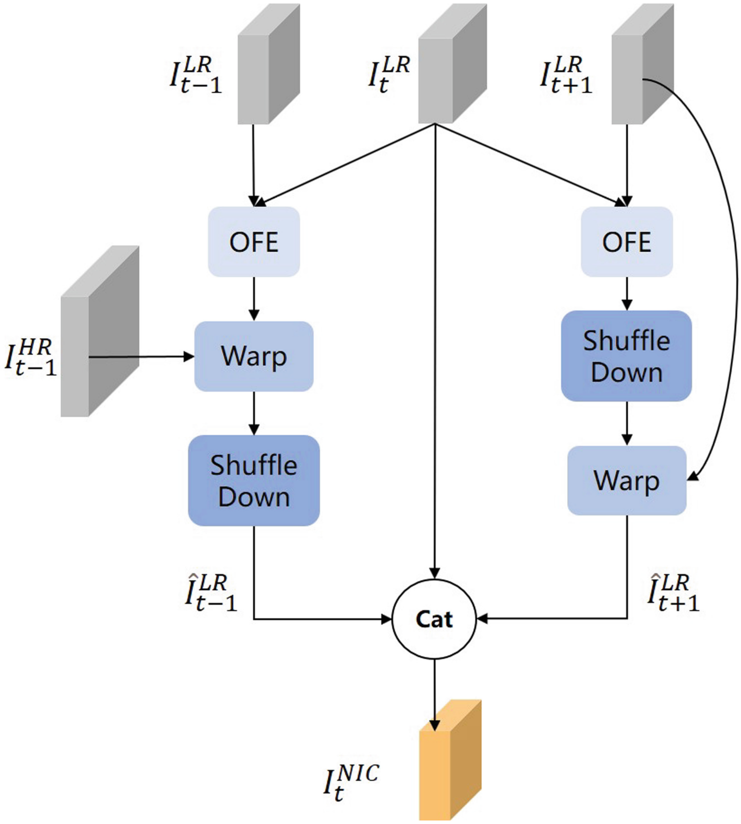

Fig. 3. Architecture of our NIC module.

$$OF^{HR}_{t+1} = OFE(I^{LR}_{t+1},\ I^{LR}_{t}) \tag{2}$$

where OF^{HR}_{t-1} and OF^{HR}_{t+1} represent the HR optical flow generated by OFE module respectively.

Shuffling is used for space-to-depth conversion, which uses the scale factor k to map LR to HR space. The operation is reversible and can achieve the inverse mapping from HR to LR. In previous work [8,16,17] shuffling has been used to change the spatial resolution of feature.

$$s^{LR} \in \mathbb{R}^{H\times W\times C} \xrightarrow{\text{shuffle up}} s^{HR} \in \mathbb{R}^{kH\times kW\times C/k^2} \tag{3}$$

$$s^{HR} \in \mathbb{R}^{H\times W\times C} \xrightarrow{\text{shuffle down}} s^{LR} \in \mathbb{R}^{H/k\times W/k\times k^2C} \tag{4}$$

For I^{LR}_{t+1}, input it together with I^{LR}_{t} into OFE module to generate HR optical flow OF^{HR}_{t+1}. After shuffling down, the LR flow cube OF^{LR}_{t+1} is generated.

$$OF^{HR}_{t+1} \in \mathbb{R}^{H\times W\times C} \xrightarrow{\text{shuffle down}} OF^{LR}_{t+1} \in \mathbb{R}^{H/k\times W/k\times k^2C} \tag{5}$$

$$\hat{I}^{LR}_{t+1} = WP(I^{LR}_{t+1}, OF^{LR}_{t+1}) \tag{6}$$

where $WP(\cdot)$ denotes warping operation and $\hat{I}^{LR}_{t+1}$ is the aligned frame at time $t+1$.

For I^{LR}_{t-1}, input it together with I^{LR}_{t} into OFRnet to generate HR optical flow OF^{HR}_{t-1}. Different from I^{LR}_{t+1}, I^{HR}_{t-1} can be fed back to the NIC module for auxiliary information fusion since it already has the HR result I^{HR}_{t-1} of the previous frame. The output feedback can improve the continuity between frames, and the generated frames are more stable. It can effectively use the information of the previous frames, which is equivalent to the accumulation of picture information in the video. Compared with I^{LR}_{t-1}, the HR frame I^{HR}_{t-1} can provide more information to help generate I^{HR}_{t}. So we warp I^{HR}_{t-1} directly using HR optical flow OF^{HR}_{t-1} to get $\hat{I}^{HR}_{t-1}$ and align it to the current frame. Finally, shuffling down $\hat{I}^{HR}_{t-1}$ to ensure consistency with $\hat{I}^{LR}_{t+1}$ and $\hat{I}^{LR}_{t}$.

$$\hat{I}^{HR}_{t-1} = WP(I^{HR}_{t-1}, OF^{HR}_{t-1}) \tag{7}$$

$$\hat{I}^{HR}_{t-1} \in \mathbb{R}^{H\times W\times C} \xrightarrow{\text{shuffle down}} \hat{I}^{LR}_{t-1} \in \mathbb{R}^{H/k\times W/k\times k^2C} \tag{8}$$

For $\hat{I}^{LR}_{t+1}$, $\hat{I}^{LR}_{t}$ and $\hat{I}^{LR}_{t-1}$, concatenating them together to get the final output I^{NIC}_{t} of the NIC module.

3.3 Face Super-Resolution Cell (FSRC)

The purpose of FSRC is to extract prior information of face, utilize the previous NIC module and UP structure, so that the facial features can flow over time and be selectively accumulated.

FSRNet [18] proposed the method of combining image features and prior information to carry out FSR. Inspired by FSRNet, we designed the following network structure (see Fig. 4) to extract prior information of face.

The network consists of FSR encoder, FSR decoder and prior information extraction (PIE) module. The input of the network is the video frames coupled with the neighborhood information and the historical hidden state. After the coupled video frame I_t^{NIC} is concatenated with the hidden state h_{t-1}, it is sent to the FSR encoder to extract image features. In another branch, the coupled video frames I_t^{NIC} are sent to the PIE module to extract facial prior information. The extracted prior information is concatenated with the image features generated by the encoder, and then sent to the FSR decoder to generate hidden state h_t and the filtered facial features I_t^{FSR}.

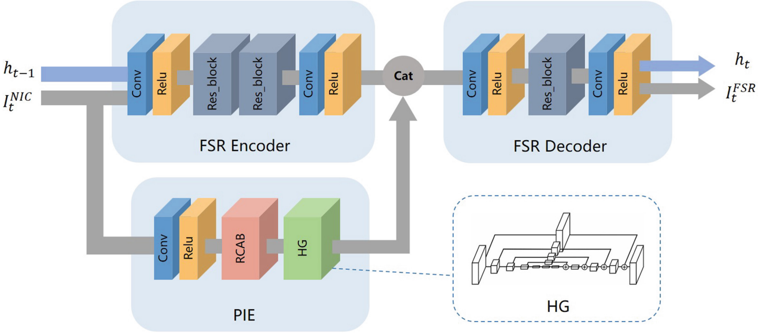

Fig. 4. Architecture of our FSRC. 'Cat' denotes concatenation along the channel dimension, 'RCAB' denotes residual channel attention block. 'HG' denotes HourGlass structure, which uses a skip connection mechanism between symmetrical layers.

Encoder-Decoder Structure. Inspired by the success of ResNet [19] in SR, we use residual blocks for feature extraction, and the network structure is shown in Fig. 4. The FSR encoder consists of convolutional layers, ReLU [20] layers and two residual blocks to extract the features of the coupled video frame I_t^{NIC} and the hidden state. The FSR decoder consists of convolutional layers, ReLU layers and a residual block, which jointly utilizes features and prior information for face image restoration.

Prior Information Extraction (PIE). In many CNN-based methods [21,22], information is treated equally in all channels during feature extraction, which makes the network lack the ability of discriminative learning. RCAN [23] proposes a deep residual channel attention network to obtain better performance. Inspired by RCAN, we add the residual channel attention block (RCAB) to re-weight the distribution of different channels.

In addition, inspired by the success of stacked heatmap regression in human pose estimation [24] and human face image SR [18], we added an HourGlass (HG) structure [24] after RCAB to effectively integrate cross-scale features and preserve spatial information at different scales. The network structure of PIE is shown in Fig. 4, consisting of convolutional layers, ReLU layers, RCAB and HG structure.

3.4 Loss Function

Adding constraints will bring more prior information to the SR process, and can effectively constrain the distribution of solutions, so as to obtain more accurate results. In the process of FVSR, we can also use the features of face to constrain the spatial distribution of solutions in a more precise way.

Due to the particularity of human face, the effective information in the face is mainly concentrated in the facial organs, so adding additional loss functions in the training process can achieve better results. In addition to the mean square error (MSE) loss $\mathcal{L}_{MSE}$ of SR frame and groundtruth (GT) frame, we add the loss of facial organs $\mathcal{L}_{face\_organ}$. For facial frames, MTCNN [25] is used to pre-calibrate the specific positions of the facial organs. Then additional MSE calculation is performed between HR frame and GT frame in corresponding facial organ regions, where $\Phi_i(i = 1, 2, 3, 4)$ represent the left eye, right eye, nose and mouth components respectively, I_t^{SR} represent HR frame of time t and I_t^H represent GT frame.

$$\mathcal{L}_{MSE} = \left\| I_t^{SR} - I_t^H \right\|_2^2 \tag{9}$$

$$\mathcal{L}_{face\_organ} = \sum\nolimits_{i=1}^{4} \left\| \Phi_i(I_t^{SR}) - \Phi_i(I_t^H) \right\|_2^2 \tag{10}$$

In addition to pixel-level differences, we can also add loss of high-level information such as image structure, texture, and style. With reference to Insightface [14], face feature extraction can be performed on face images to generate a vector containing face identity information. The loss function $\mathcal{L}_{face\_vector}$ is constructed according to the Euclidean distance of the feature vector between HR frame and GT frame, where Θ represents using Insightface to extract face features. According to this training method, more accurate recognition results can be obtained after FSR.

$$\mathcal{L}_{face\_vector} = \left\| \Theta(I_t^{SR}) - \Theta(I_t^H) \right\|_2^2 \tag{11}$$

Based on the above analysis, we design three loss terms, the MSE loss of HR frame and GT frame $\mathcal{L}_{MSE}$, and the MSE loss of the pixels in facial features area $\mathcal{L}_{face\_organ}$ and the loss of high-level feature vectors of faces $\mathcal{L}_{face\_vector}$.

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda_1 \mathcal{L}_{face\_organ} + \lambda_2 \mathcal{L}_{face\_vector} \tag{12}$$

4 Experiments

In this section, we first compare our framework to several existing VSR methods. Then, we further conduct ablation experiments to evaluate our framework.

4.1 Dataset

Due to the lack of recognized FVSR Dataset, we make a Face Video Dataset made by 300 Videos in the Wild (300-VW) and conducted experiments on it. It was first used in ICCV's face recognition contest in 2015 and can be downloaded at ibug.doc.ic.ac.uk. According to requirements, our dataset production process is as follows. First, the original videos are intercepted into sequences of consecutive frames, 32 frames per sequence, and a total of 400 video sequences are generated. MTCNN [25] network is used to select and cut facial area of each frame, and then adjust the size to 160*160 to obtain HR frames. Then we performed downsampling to generate LR frames and obtain final Face Video Dataset. In this work, we only focus on the downsampling factor of 4 since it is the most challenging case.

Finally, We divide 400 generated video sequences into training sets, verification sets and testing sets (see Table 1).

Table 1. Datasets used in FVSR.

Face video dataset	Sequences	Frames
Training	340	10880
Validation	15	480
Testing	45	1440

4.2 Implementation Details

To train our network, we randomly selected 10 consecutive frames from 32 frames. Due to UP structure, the hidden state h_{t-1} and the HR result I_{t-1}^{HR} of the previous frame need to be initialized. Both tensors are initialized with zeros.

We implemented our framework in Pytorch. We set the batch size to 4, the learning rate to 10^{-4}, and λ_1 and λ_2 to 0.05 and 0.01, respectively. All experiments are conducted on a PC with an Nvidia GTX 1080Ti GPU.

4.3 Comparisons to the State-of-the-Art

Quantitative Comparisons. We compare our proposed FAPN with the state-of-the-art VSR methods, including VESPCN [9], FRVSR [16], SOF-VSR [15] and

Table 2. Quantitative comparisons. Best results are shown in boldface.

	SOF-VSR [15]	FRVSR [16]	TecoGan [26]	VESPCN [9]	Ours
PSNR (dB)	30.75	25.25	24.85	30.00	**31.06**
SSIM (dB)	0.917	0.823	0.811	0.907	**0.924**
Face distance	0.201	0.276	0.252	0.228	**0.197**

TecoGan [26]. Quantitative comparison with other state-of-the-art VSR methods is shown in Table 2.

We conducte experiments on 45 test sets and measure PSNR, SSIM and Face distance, which measures the difference between faces. Face_recognition [13] is a concise and powerful face recognition library, tested with the Labeled Faces in the Wild face dataset, with an accuracy rate of 99.38%. We use it to measure the Face distance metric. Face_recognition generates high-dimensional feature vectors for face images, and then calculates the Euclidean distance between corresponding face feature vectors of HR frames and GT frames to quantify the differences between faces. The smaller the Face distance is, the higher the face similarity is.

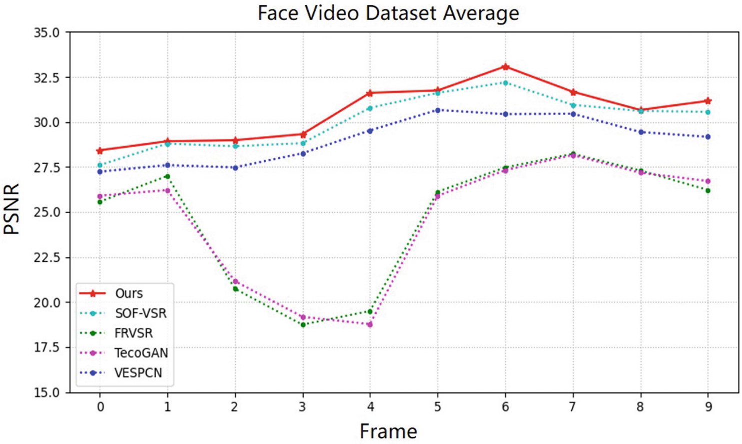

Fig. 5. Average PSNR values for the first 10 frames on the test sets. Average PSNR values of each frame is calculated from the average of the corresponding frames on all test sets

The experimental results demonstrated that the proposed framework FAPN achieves better performance on PSNR, SSIM and Face distance compared with state-of-the-art methods. Specifically, the PSNR and SSIM values achieved by our framework are better than other methods by over 0.31 dB and 0.15 dB.

This is because we use NIC to supplement and fuse information, and the information provided to the UP structure is more abundant and accurate, therefore, more reliable spatial details and temporal consistency can be recovered well.

For Face distance, we achieved an improvement of 1.99%, which indicates that the face generated by our network is closer to real face and has high reliability.

It can be seen that the PSNR and SSIM values of the TecoGan [26] and FRVSR [16] networks are only about 25 dB and 0.8 dB, and the Face distance value exceeds 0.25. This is due to the non-natural generation of SR frames, with severe facial deformation or incorrect information generation. The specific test images will be displayed in the Qualitative comparisons. In addition, we plot the average PSNR for the first 10 frames on the test sets. As can be seen from Fig. 5, our network achieves the best PSNR values for each frame.

Qualitative Comparisons. A qualitative comparison between our method and other SR methods [9, 15, 16, 26] are shown in Fig. 6. There are three face images, each of which is reconstructed from ten consecutive frames. The results obtained by our network are closer to real frames than other methods, especially in facial features.

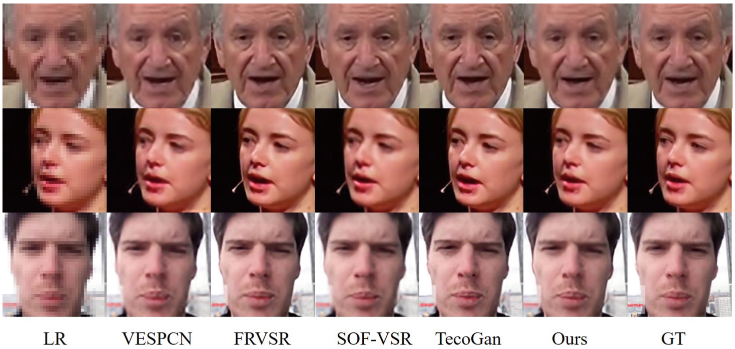

Fig. 6. Qualitative comparisons, where all the examples come from our test sets.

It can be seen that the visual effect represented by TecoGan [26] is better, but their PSNR, SSIM and Face distance metrics are poor. Although they generate richer details, they are generated incorrectly, resulting in a certain degree of distortion of human facial features. For example, the wrinkles on the forehead of the old man in the first row are falsely generated. In addition, the mouth of the little girl generated by TecoGan [26] and FRVSR [16] in the second row is deformed. These additionally generated textures and deformations will make the visual effect better to a certain extent, but will reduce the authenticity of the generated face. From the perspective of the actual application scenario of FVSR, it is expected to improve the accuracy of face recognition through super-resolution. Therefore, it is necessary to include more details as much as possible on the premise of ensuring the authenticity of the generated face. So compared with TecoGan [26] and FRVSR [16], our results are closer to the real situation.

From the 32 frames of the test sets, we selected 6 consecutive frames for testing, and the results are shown in Fig. 7. Expanding horizontally in chronological order, it can be seen that for TecoGan [26] and FRVSR [16], the faces in columns 4 and 5 are severely distorted. This is because when the inter-frame motion is intense to a certain extent, inaccurate motion estimation and compensation not only cannot effectively utilize the inter-frame information, but also introduce error information and seriously interfere with SR results. Our network utilizes NIC module, which can not only effectively utilize the information of previous frame to supplement the information of the current frame, but also ensure the authenticity of the introduced information to avoid unnatural generation.

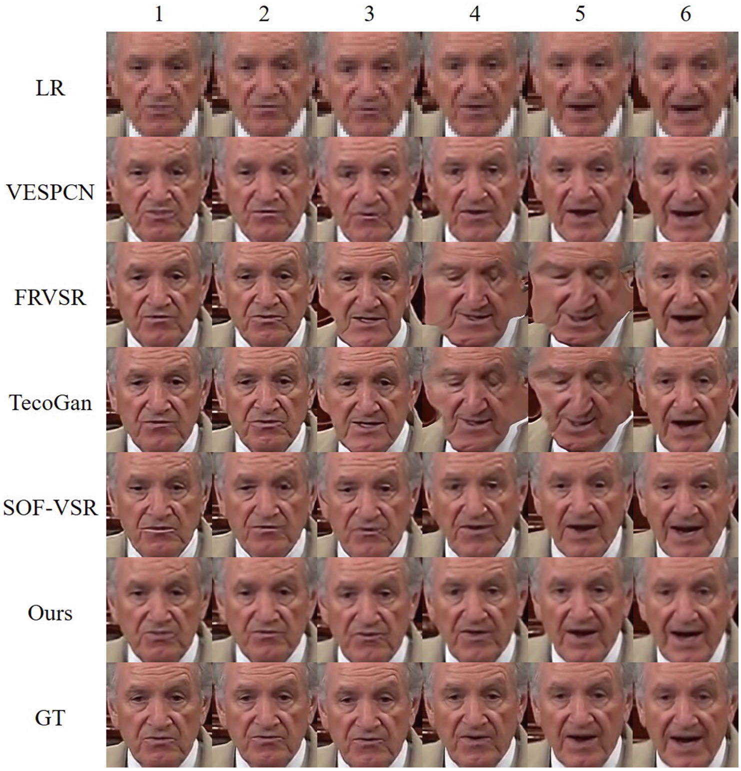

Fig. 7. Qualitative comparisons of 6 consecutive test frames.

Due to UP structure, our network relies on historical input frames. Initially, the information content available is minimal, and a certain number of input frames are required to accumulate information. This phenomenon can be seen in Fig. 7. In the first two frames, the texture on the forehead of the old man, and the eyes have a certain degree of distortion, and the facial structure of the human face cannot be completely reconstructed. But with more information, the textures on the facial features and forehead gradually approach the real face.

In addition, compared with SOF-VSR [15] and VESPCN [9], we slightly improve visual effects and indicators, introducing richer details on the premise of ensuring authenticity.

4.4 Ablation Study

Effects of NIC and FSR. We conducted 3 experiments to evaluate the effects of NIC and FSR, respectively. Specifically, we remove the NIC and FSRC from our network, the remaining parts constitute the first network, named 'BasicNet v1'. The second network, named 'BasicNet v2', has the same structure as FAPN except for NIC module. In this part, we study the effects of different networks. For fairly comparison, we train all those models with other same implementation details.

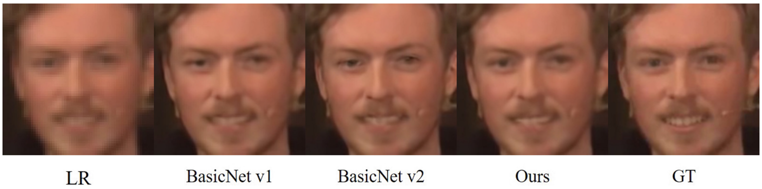

Fig. 8. Qualitative results of ablation study.

For 'BasicNet v1', as Fig. 8 shows, the generated face is blurry, and the left eye is distorted relative to GT. For 'BasicNet v2', the face is clear, but there exist distortions and false generation. This is because 'BasicNet v2' introduced FSRC compared to 'BasicNet v1', which can advance the high-level information of the face, but without NIC, it may lead to the accumulation of wrong information. The results of FAPN are closest to GT, achieving correct and sufficient information extraction. The results of the quantitative comparison are shown in Table 3. PSNR, SSIM and Face distance of 'BasicNet v2' are the worst, which indicates that we should not only focus on visual effects, but should take authenticity as an important evaluation factor. Compared to 'BasicNet v1', our final network has 0.68 dB improvement in PSNR, 0.004 dB improvement in SSIM, and 0.005 improvement in Face distance.

Effects of Loss Function. In this section, we adopt 4 training methods, and evaluated the results respectively for our PAFN network. Where 'MSE' stands for MSE loss. 'MSE+GAN' refers to adding an adversarial network on the basis of MSE, and forms an adversarial loss [27] with the help of the generator and

Table 3. Quantitative results of ablation study. Best results are shown in boldface.

	BasicNet v1	BasicNet v2	FAPN
PSNR (dB)	29.58	29.37	**30.26**
SSIM (dB)	0.918	0.823	**0.922**
Face distance	0.197	0.206	**0.192**

the discriminator. 'MSE+VGG' refers to using the VGG network for high-level feature extraction of the face based on the MSE and drawing on the powerful feature extraction characteristics of the VGGNet [28]. The loss function includes not only the MSE loss, but also the mean squared loss of each layer of feature maps of the VGG feature extraction module, in order to obtain the similarity between high-level features of the image. 'MSE+FACE' refers to the loss function we constructed in Sect. 3.5. We omit the different loss weight adjustment steps for each method, and the results of each method under the optimal parameters are shown in Table 4.

Table 4. Ablation study on the effects of different loss terms. Best results are shown in boldface.

	MSE	MSE+GAN	MSE+VGG	MSE+FACE
PSNR (dB)	30.88	30.92	30.50	**31.00**
SSIM (dB)	0.912	**0.913**	0.902	**0.913**
Face distance	0.215	0.206	0.214	**0.203**

It can be seen that 'MSE+FACE' achieves the best performance on PSNR (up to 0.08 dB), SSIM (same as 'MSE+GAN') and face recognition accuracy (up to 1.46%), because it not only emphasizes the facial features, but also uses the Insightface [14] network to extract the high-level features of the face.

5 Conclusion

In this paper, we propose an end-to-end face alignment propagation network (FAPN) for face video super-resolution. Our NIC module first fuses adjacent frames and previous HR frame. UP structure and FSRC are then performed to accumulate correct facial prior information. Extensive experiments have demonstrated that our FAPN can recover facial details and improve the accuracy of face recognition on the premise of ensuring the authenticity of the generated face. Comparison to existing video SR methods has shown that our framework achieves the state-of-the-art performance on PSNR (up to 0.31 dB), SSIM (up to 0.15 dB) and face recognition accuracy (up to 1.99%).

References

1. Wang, M., Deng, W.: Deep face recognition: a survey. Neurocomputing **429**, 215–244 (2021)
2. Farooq, M., Dailey, M., Mahmood, A., Moonrinta, J., Ekpanyapong, M.: Human face super-resolution on poor quality surveillance video footage. Neural Comput. Appl. **33**, 13505–13523 (2021)
3. Yu, F., Li, H., Bian, S., Tang, Y.: An efficient network design for face video super-resolution. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1513–1520 (2021)
4. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3892–3901 (2019)
5. Wang, X., Chan, K.C., Yu, K., Dong, C., Loy, C.C.: EDVR: video restoration with enhanced deformable convolutional networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1954–1963 (2019)
6. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3224–3232 (2018)
7. Isobe, T., Jia, X., Gu, S., Li, S., Wang, S., Tian, Q.: Video super-resolution with recurrent structure-detail network. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 645–660. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_38
8. Fuoli, D., Gu, S., Timofte, R.: Efficient video super-resolution through recurrent latent space propagation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3476–3485 (2019)
9. Caballero, J., et al.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2848–2857 (2017)
10. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: BasicVSR: the search for essential components in video super-resolution and beyond. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4945–4954 (2021)
11. Xin, J., Wang, N., Li, J., Gao, X., Li, Z.: Video face super-resolution with motion-adaptive feedback cell. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12468–12475 (2020)
12. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: a general-purpose face recognition library with mobile applications (2016)
13. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4685–4694 (2019)
15. Wang, L., Guo, Y., Lin, Z., Deng, X., An, W.: Learning for video super-resolution through HR optical flow estimation. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11361, pp. 514–529. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20887-5_32

16. Sajjadi, M.S.M., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6626–6634 (2018)
17. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1874–1883 (2016)
18. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: FSRNet: end-to-end learning face super-resolution with facial priors. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2492–2501 (2018)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
20. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines vinod nair. In: International Conference on International Conference on Machine Learning (2010)
21. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25
22. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1132–1140 (2017)
23. Basak, H., Kundu, R., Agarwal, A., Giri, S.: Single image super-resolution using residual channel attention network. In: 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), pp. 219–224 (2020)
24. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
25. Zhang, L., Wang, H., Chen, Z.: A multi-task cascaded algorithm with optimized convolution neural network for face detection. In: 2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS), pp. 242–245 (2021)
26. Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thuerey, N.: Learning temporal coherence via self-supervision for GAN-based video generation. ACM Trans. Graph. (TOG) **39** (2020)
27. Goodfellow, I.J., et al.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS 2014, pp. 2672–2680. MIT Press, Cambridge (2014)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2015)