第51卷第10期 中国科学技术大学学报 Vol. 51, No. 10
2021年10月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Oct. 2021
Received: 2020-12-24; Revised: 2021-02-05 doi:10.52396/JUST-2020-0032

A low-latency inpainting method for unstably transmitted videos

WEI Yutong[1], BAO Bingkun[2], ZHANG Ziqi[3], ZHU Jin[1]*
1. Department of Automation, University of Science and Technology of China, Hefei 230027, China;
2. College of Telecommunications Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
3. Agora, Inc., Shanghai 200082, China
* Corresponding author. E-mail: jinzhu@ustc.edu.cn

Abstract: Video traffic has gradually occupied the majority of mobile traffic, and video damage in unstable transmission remains a common and urgent problem. The difficulty of inpainting these damaged videos is that the holes randomly appear in random video frames, which are hard to be well settled with both low latency and high accuracy. We are the pioneer to look into the video inpainting task in unstable transmission and propose a low-latency video inpainting method which consists of two stages: In the coarsely inpainting stage, we achieve the extraction of damaged two-dimensional optical flow from reference frames, and establish a linear prediction model to coarsely inpaint the damaged frames according to the temporal consistency of motions. In the fine inpainting stage, a Partial Convolutional Frame Completion network (PCFC-Net) is proposed to synthesize all reference information and calculate a fine inpainting result. Compared with that of the state-of-the-art baselines, the waiting time for reference frames is greatly reduced while PSNR and SSIM are improved by 4.0% ~12.7% on DAVIS dataset.
Keywords: video inpainting; unstable transmission; partial CNN; linear prediction
CLC number: TP273 **Document code**: A

1 Introduction

According to the white paper of visual networking index released by Cisco[1], video data has already accounting for more than half of all mobile data traffic. However, the problem of video damage due to the unstable transmission has not been solved yet, which affects the audience perception greatly. For this reason, improving the quality of the transmitted video becomes critical and we focus on the inpainting of damaged videos.

Among the wide range of flexible implementation of video inpainting, lots of efforts can nowadays be referred to, which can be roughly divided into two categories: the traditional mathematical method and the deep learning method. The former searches for the useful parts from the reference frames, and copying or merging them to inpaint the damaged frames, e. g. Refs. [2-4]. Although this method achieves varying degrees of success, the high calculation cost and low learning ability for deep information still prevent it from application. While the latter focuses on using the deep neural network to extract both temporal and spatial information from the reference frames for filling the holes. Wang et al[5] proposed a deep learning structure with a 3DCN and a CombCN, which directly extracts useful information from RGB reference frames to inpaint the damaged RGB frame. Xu et al[6] considered the inpainting task as a pixel propagation problem, and introduced the optical flows to simplify the inpainting task. These two works achieve state-of-the-art performance, but require the complete video clip to provide enough reference information, which brings a relatively long waiting period for the collection of reference frames in real-time video transmission. By combing both RGB reference frames and optical flows, Kim et al[7] proposed a VINet model where the length of reference frames is reduced to 6th frames before and after the damaged frame. However, this still cannot satisfy the high real-time requirements in video transmission. In the scene of video transmission, an effective model should minimize the waiting time of the reference frame while ensuring high video quality.

To achieve the above mentioned two goals, a low-latency inpainting method is proposed to consist of two stages: coarse inpainting and fine inpainting. In the coarse inpainting stage, we use masks to enable FlowNet 2.0 to extract the optical flows between these reference frames. Then according to the temporal consistency of motions, we can obtain the estimated optical flows between reference frames and the damaged

Citation: Wei Yutong, Bao Bingkun, Zhang Ziqi, et al. A low-latency inpainting method for unstably transmitted videos. J. Univ. Sci. Tech. China, 2021, 51(10): 717-724, 786.

frame via a linear combination of reference flows, which is used to warp reference frames into coarse inpainting results. Due to the fact that the correlation between the reference frames and damaged frame is actually more complicated, the coarse inpainting results will have obvious artifacts compared with the ground-truth. This drives us to design a fine inpainting stage in order to improve the quality of these coarse inpainting results in which these coarse inpainting results can take the role of additional reference information. In the fine inpainting stage, a partial convolutional frame completion network (PCFC-Net) is proposed to synthesize all reference information and calculate a fine inpainting result. With the help of powerful deep information extraction and synthesis capabilities of the deep network, the inpainting effect of damaged frames can be greatly improved.

Compared with the state-of-the-art baseline[7], the number of reference frames required is reduced by 2/3 compared to the original. And the experimental results show that the PSNR and SSIM of our method are both improved to a relatively higher level than the state-of-the-art baseline.

2 Video inpainting task under unstable transmission

The inpainting task under unstable transmission is different from the well-studied inpainting tasks such as object removal, as a pioneer in this research direction, we formulate the inpainting task for video damage in unstable transmission as follows: the critical feature of video damage is completely random, as shown in Figure 1, which consists of the following five features: ① Random frames. The holes caused by transmission may appear on any frames of the video, including the reference frames, which requires the inpainting model to have the ability to exploit incomplete reference frames for inpainting the damaged frame. ② Random positions. The position of holes on a certain frame is also random. When the holes on two adjacent frames appear in different positions, the deviation of the inpainting results of the two adjacent frames will be enlarged by comparison, so the inpainted content must be consistent with the the real content for the video coherence, otherwise the inpainting deviation will not only affect the quality of the current frame, but also affect the continuity of the video content. ③ Random number. For the reason that the packet loss is completely uncontrollable, there may be more than one hole on a certain frame. In extreme cases, all content of the frame is lost. ④ Random size. Due to different transmitting strategies, the size of holes caused by packet loss is different. These two characteristics cause us to inevitably face the difficult situation of repairing large-scale holes, where the residual information on the damaged frame and incomplete reference frames becomes highly insufficient. ⑤ Random shape. The original video before transmission may have flaws, appearing in irregular shapes, which causes additional challenges to inpainting tasks.

Figure 1. Video damage in unstable transmission

For such general inpainting task, the proposed method must achieve two aspects of advantages: firstly the reference frames sampling window should be as small as possible for the sake of reducing latency. And secondly the inpainting effects, such as video coherence and scene restoration, should be superior to the existing

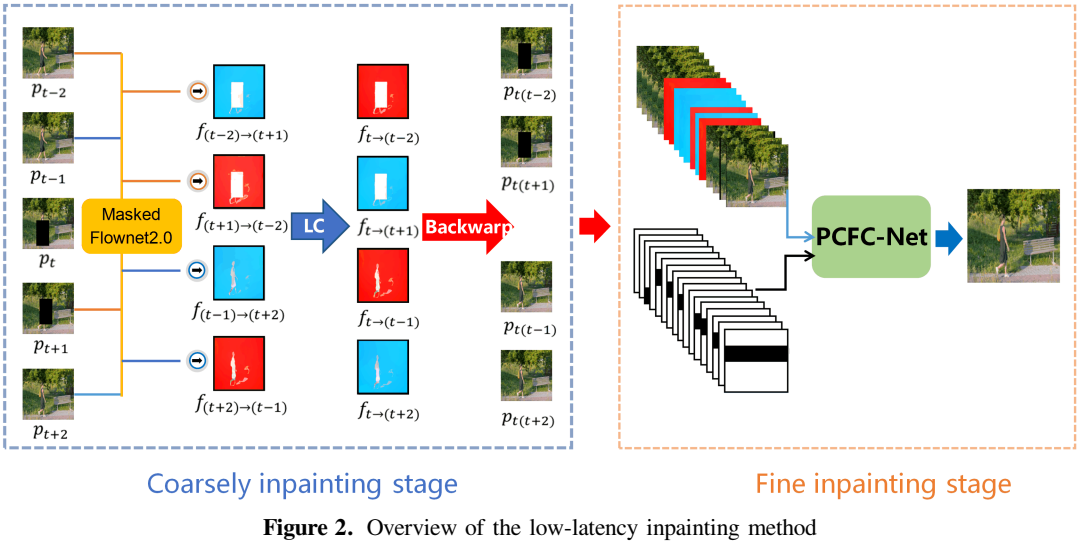


Figure 2. Overview of the low-latency inpainting method

achievements simultaneously.

3 Low-latency inpainting method

Let the size of the sampling window be n, then there are n reference frames needed for the inpainting the damaged frame P_t: n frames $P_{t-n}\cdots P_{t-1}$ before and n frames $P_{t+1}\cdots P_{t+n}$ after P_t. Since $P_{t-n}\cdots P_{t-1}$ are in front of P_t, they will be inpainted before P_t even if they are damaged. In opposite, when P_t is under inpainting, $P_{t+1}\cdots P_{t+n}$ will be used as reference frames directly whether they are damaged or not. We set $n=2$ out of the consideration of saving waiting time and computation cost, and propose a low-latency inpainting method to reconstruct the damaged video progressively as shown in Figure 2, which can be introduced in two parts:

3.1 Coarsely inpainting stage based on linear combination

In order to differ from the existing achievements which usually utilizes the temporal information hidden in RGB images directly, We extract the two-dimensional optical flow between the reference frames in order to learn and reconstruct the motion trajectory between frames. The introduction of optical flows, which are used to represent the movement speed and direction of each pixel in two adjacent frames, significantly reduces the difficulty of extracting temporal information and finally leads to the simplification of inpainting tasks.

We adopt the state-of-the-art model of optical flow extraction-Flow Net 2.0[8] to extract the optical flows $F_{(t-2)\to(t+1)}$, $F_{(t+1)\to(t-2)}$ between P_{t-2}, P_{t+1} and $F_{(t-1)\to(t+2)}$, $F_{(t+2)\to(t-1)}$ between P_{t-1}, P_{t+2}. However, due to the complete randomness of video damage, the reference frames P_{t+1}, P_{t+2} are often with holes and this model may not take effect because the pixels in the holes can not be tracked by FlowNet 2.0 when P_{t+1}, P_{t+2} are incomplete, as shown in Figure 3.

In order to avoid this situation mentioned above, we modify the inputs of FlowNet 2.0 where masks are used to represent the holes in of P_{t+1}, P_{t+2}. And we put the same masks on P_{t-2}, P_{t-1} respectively.

$$M_i[x,y]=\begin{cases}0, & \text{if pixel } (x,y) \text{ in frame i is damaged}\\ 1, & \text{otherwise}\end{cases} \tag{1}$$

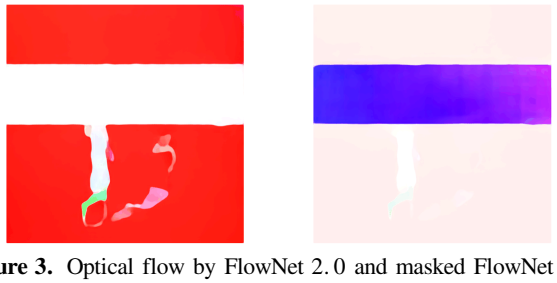


Figure 3. Optical flow by FlowNet 2.0 and masked FlowNet 2.0

where M_i represents mask of frame i.

After the masking operation, we can obtain the optical flows of incomplete frames with Equation (2), as shown in Figure 3(b).

$$\begin{aligned}F_{i\to j}&=\Psi(P_i\odot M_j,\ P_j\odot M_j),\\ F_{j\to i}&=\Psi(P_j\odot M_j,\ P_i\odot M_j),\\ i&=t-2,t-1;j=t+1,t+2\end{aligned} \tag{2}$$

where Ψ denotes FlowNet 2.0, and $\odot$ denotes element-wise multiplication.

With these masked optical flows, the preliminary inpainting process can be performed by reconstructing of the optical flows between P_t and four reference frames, which accords to the temporal continuity of the motion.

Inspired by Reference [9] and considering the smoothness of the movement, we can then calculate the estimated flows between the reference frames and the damaged frame according to the relative distance of the damaged frame and the reference frames on the time axis.

$$\begin{aligned}\widehat{F}_{t\to j}&=(1-\mathrm{e}^{\frac{t-j}{t-i}})F_{i\to j}=-(1-\mathrm{e}^{\frac{t-j}{t-i}})F_{j\to i},\\ \widehat{F}_{t\to i}&=\mathrm{e}^{\frac{t-j}{t-i}}F_{j\to i}=-\mathrm{e}^{\frac{t-j}{t-i}}F_{i\to j},\\ i&=t-2,t-1;\ j=t+1,t+2\end{aligned} \tag{3}$$

where $\widehat{F}_{t\to(t-1)}$, $\widehat{F}_{t\to(t-2)}$, $\widehat{F}_{t\to(t+1)}$, $\widehat{F}_{t\to(t+2)}$ denote the estimated flows, as shown in Figure 4 and $\mathrm{e}^{\frac{i-t}{i-j}}$ is a weight factor which represents the distance between the damaged frame P_t and the reference frames P_i, P_j in the time dimension.

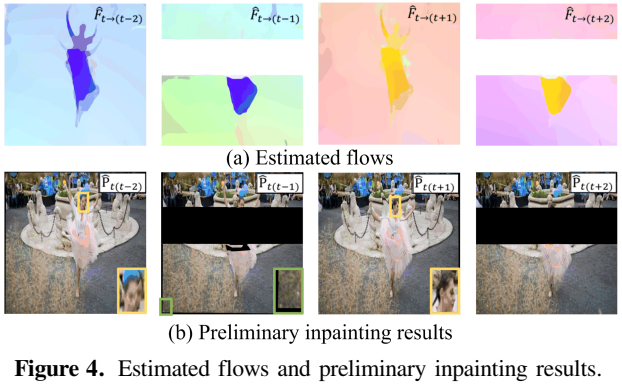
Rewriting Equation (3) we can get the calculation formula of the estimated optical flows:

$$\begin{aligned}\widehat{F}_{t\to j}&=(1-\mathrm{e}^{\frac{t-j}{t-i}})^2F_{i\to j}-(1-\mathrm{e}^{\frac{t-j}{t-i}})\mathrm{e}^{\frac{t-j}{t-i}}F_{j\to i},\\ \widehat{F}_{t\to i}&=-(1-\mathrm{e}^{\frac{t-j}{t-i}})\mathrm{e}^{\frac{t-j}{t-i}}F_{i\to j}+(\mathrm{e}^{\frac{t-j}{t-i}})^2F_{j\to i},\\ i&=t-2,t-1;\ j=t+1,t+2\end{aligned} \tag{4}$$

Then using these estimated flows, we can warp four reference frames to obtain the preliminary inpainting results of P_t.

$$\begin{aligned}\widehat{P}_{ti}&=W(\widehat{F}_{t\to i},\ P_i),\\ i&=t-2,\ t-1,\ t+1,\ t+2\end{aligned} \tag{5}$$

where $\widehat{P}_{ti}$ represents the preliminary inpainting results of



(a) Estimated flows

(b) Preliminary inpainting results

Figure 4. Estimated flows and preliminary inpainting results.

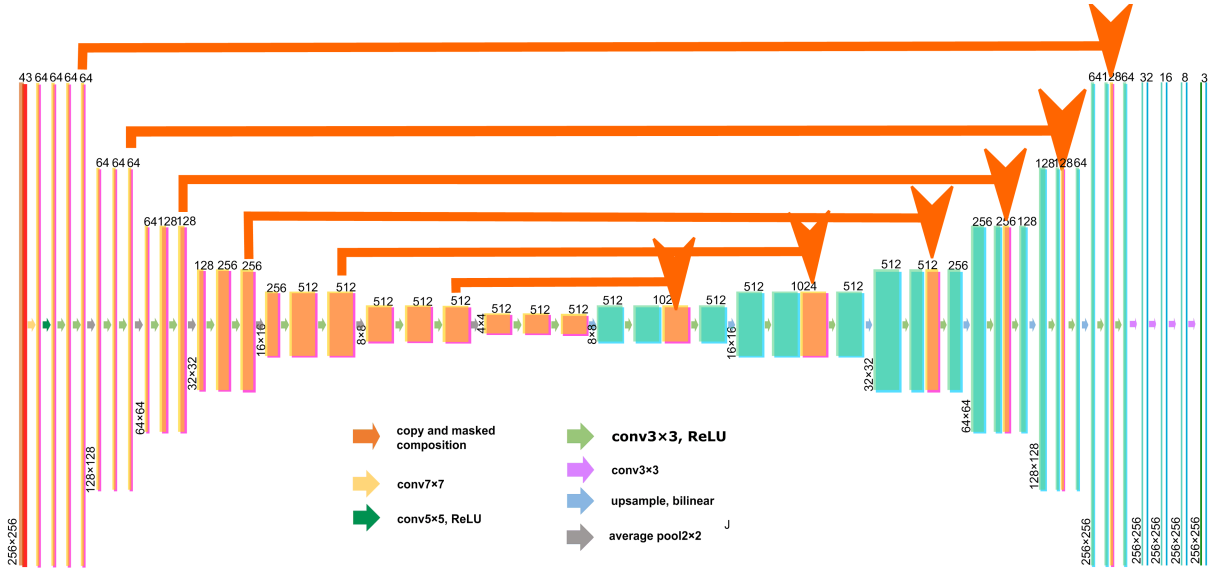


Figure 5. PCFC-Net. The yellow boxes represent the frames of input (darker one) and the corresponding feature maps (lighter ones) in the encoder. The pink boxes represent the corresponding masks of frames and feature maps in the encoder. The green boxes represent the feature maps (lighter ones) in the decoder and the output picture (darker one) of the model. And the blue boxes represent the corresponding masks of the feature maps and the output picture.

P_t and $W(\cdot,\cdot)$ denotes the backward warping function. Applying this warping operation on the estimated flows and reference frames, we can get the preliminary inpainting results, as shown in Figure 4 (b).

It can be seen from Figure 4 that part of the preliminary results remains unknown due to the incomplete estimated optical flows, meanwhile these four preliminary results are different which is caused by the estimated bias of estimated flows (as shown in the yellow box), and for the reason that the estimated optical flows cannot track every pixel of P_t, black edges appear (as shown in the blue box). The inpainting results with such obvious artifacts certainly cannot satisfy the high requirements of video coherence and scene restoration in the video transmission scene, which drives us to propose a neural network to synthesize all these four useful preliminary results and generate more satisfying inpainting result of P_t.

3.2 Fine inpainting stage based on PCFC-net

Considering that the coarsely inpainting is only a preliminary excavation of reference information and coarsely inpainting results can not achieve the high-quality video restoration, the PCFC-Net is proposed to extract hidden information from the damaged frame (P_t), reference frames (P_{t-2}, P_{t-1}, P_{t+1}, P_{t+2}), reference flows ($F_{(t-2)\to(t+1)}$, $F_{(t+1)\to(t-2)}$, $F_{(t-1)\to(t+2)}$, $F_{(t+2)\to(t-1)}$), estimated flows ($\widehat{F}_{t\to(t-2)}$, $\widehat{F}_{t\to(t-1)}$, $\widehat{F}_{t\to(t+1)}$, $\widehat{F}_{t\to(t+2)}$) and coarsely inpainting results ($\widehat{P}_{t(t-2)}$, $\widehat{P}_{t(t-1)}$, $\widehat{P}_{t(t+1)}$, $\widehat{P}_{t(t+2)}$) in order to generate the final inpainting result, which is composed of many partial convolutional layers for the purpose of making full use of the incomplete reference frames. And this network is constructed as an encoder-decoder architecture network, as shown in Figure 5.

The PCFC-Net consists of an encoder and a decoder, corresponding to the left part and right part of Figure 5 respectively. The encoder is composed of one 7 × 7 partial convolutional layer, one 5 × 5 partial convolutional layer, two 3×3 partial convolutional layers and six duplicate of downsampling units which are made up with 2 × 2 average pooling layer, 3 × 3 partial convolutional layer and Leaky ReLU ($\alpha = 0.1$) layer in order. The average pooling layer in downsampling units is responsible for reducing picture size by 1/2. The decoder consists of six upsampling units which are made up of bilinear upsampling layer, 3 × 3 partial convolutional layer and Leaky ReLU ($\alpha = 0.1$) layer in order and four 3 × 3 partial convolutional layers. The bilinear upsampling layer in upsampling units is responsible for restoring the original size of the pictures. By each upsampling, the picture size doubles. And the feature maps from the downsampling units are passed to the upsampling units with the same resolution, as shown by the yellow connecting line in Figure 5, the undamaged part of which provides reference information for upsampling and helps to improve the quality of restored pictures, but the feature maps passed by these connections may also bring 0 values in the masked area, as Reference[7] concerns. So we linearly combine two feature maps in the encoder and decoder to replace the feature maps passed from the encoder.

$$S_i = E_i \odot M_{E_i} + D_i \odot (1 - M_{E_i}) \quad (6)$$

where E_i denotes the feature map passed by the skip connection from the encoder, M_{E_i} denotes the corresponding mask, and D_i denotes the feature map in the decoder. Replacing E_i with S_i can avoid the issue mentioned above.

In order to obtain the best training effect, after analyzing and experimenting, we used a combination of multiple loss functions to train the PCFC-Net.

Our ultimate goal is to train the PCFC-Net to have the ability to produce inpainting results exactly the same as the ground-truth, so we choose l_1-distance to compare the inpainted image P_{predict} and the corresponding ground-truth P_{t_gt} pixel by pixel.

$$L_1 = \| P_{\text{predict}} - P_{t\_gt} \|_1 \quad (7)$$

The l_1-loss function forces the output to be exactly the same with the ground-truth, but the inpainting result is often blurry due to the shortcomings of CNN models in recovering fine textures. To help solve this problem, we adopt the perceptual loss which can encourage the output to be more semantically similar to the ground-truth[9].

$$L_2 = \frac{1}{\text{CHW}} \| \phi(P_{\text{predict}}) - \phi(P_{t\_gt}) \|_2^2 \quad (8)$$

where $\phi(\cdot)$ denotes the GoogLeNet[10] without fully connected layer, and CHW denotes the size of the feature map generated by GoogLeNet.

Since there are obviously more samples in the smooth region than in the boundary region, the training process is usually dominated by the smooth area, but the final inpainting result on the boundary region usually has a significant impact on the human visual effect, so we use hard flow example mining loss to make the model focus more on the boundary region.

$$L_3 = \frac{\| M_{\text{hard}} \odot (P_{\text{predict}} - P_{t\_gt}) \|_1}{\| M_{\text{hard}} \|_1} \quad (9)$$

where M_{hard} is a mask indicating the region which is hard for the model to learn to inpaint. If the loss of a certain pixel of output is in the top 50%, the pixel value at the corresponding position in M_{hard} is 1. Through this mechanism, this loss function pay more attention on the regions difficult to learn to inpaint, and these regions are often the boundary region, which is confirmed in Reference [6].

In this paper, we choose a linear combination of these three different losses with weights to get the total loss.

$$L_{\text{total}} = \alpha_1 * L_1 + \alpha_2 * L_2 + \alpha_3 * L_3 \quad (10)$$

4 Experiments

4.1 Datasets and data preparation

We adopt the YouTube-VOS dataset[11] to train our model, the train and test datasets of YouTube-VOS is used for training and the validation dataset of YouTube-VOS is used for validation. We divide one video into many groups where each group contains twelve frames (0−11), and randomly select one from the frame 2th-9th as the ground-truth P_{t_gt}, the two frames before and after P_{t_gt} are selected as reference frames consequently, then the holes with size of 1/4 of the frames are randomly added on P_{t_gt} to produce the damaged frame P_t. All frames are resized to 300×520. And to further increase the diversity of our training dataset, we randomly capture a 256×256 square of the frames and then randomly make a rotation to them.

To test the generalization ability of the model, we use the DAVIS dataset[12] for testing. Different sizes and different proportions of holes are randomly added to the video frames. And the random feature of the holes may result in extreme cases where several holes may appear consecutively on several adjacent frames or appear densely on the same frame.

All masks needed are generated according to the location where holes appear.

4.2 Training settings

We use the masked FlowNet 2.0 to generate the needed flows in the training of PCFC-Net where the weights of the total loss are empirically given as $\alpha_1 = 1, \alpha_2 = 500, \alpha_3 = 0.05$. We train the PCFC-Net for 60 epochs in total, and the optimizer is Adam with the learning rate initialized to 0.0001, which decreases by a factor of 10 every 20 epochs.

4.3 Comparisons with existing methods

4.3.1 Evaluations

We use two recognized image quality evaluation index to measure the similarity between the inpainting result and the ground-truth: Peak signal-to-noise ratio (PSNR) score and structural similarity index (SSIM) score. And we compare the proposed method with the state-of-the-art baseline of video inpainting[7] by running their model with the same experiment settings.

4.3.2 Testing Settings

Aiming at simulating different situations of video damage under different transmission strategies and different transmission conditions, we add different sizes and different proportions of holes to the frames of videos. And it should be noticed that it is possible for more than one holes appearing on the same frame, so we cannot simply use the proportion of the number of damaged frames to the number of all frames for representing the degree of video damage. Instead, in order to be more precise, we use the proportion of the total area of holes to the total area of frames to represent the degree of video damage.

We compare the performance of our method with the baseline under different sizes which are set to be 1/4, 1/6, 1/8 of the size of frames and different

proportions of holes which goes from 5% to 30%. Among them, there are always about 1% holes to present the damage of original video before transmission, which is added artificially with random shapes. When inpainting this type of hole, we can generate a mask by making the minimum bounding rectangle of the hole, so that the inpainting of this type of hole can be unified with the hole caused by an unstable transmission. When the proportion of holes is 5% and the size of holes is 1/4 of the size of frames, the proportion of damaged frames to the total number of frames is only about 20%. While when the proportion of holes reaches 30%, no matter the size of the holes, almost all frames are damaged frames. So the different combinations of these proportions and sizes can represent various transmission conditions, even the extremely poor conditions.

4.4 Effectiveness of PCFC-net

For an intuitive understanding of the impact of PCFC-Net, we compare the coarsely inpainting results with the final result generated by the PCFC-Net as shown in Table 1. And for effective comparison, we select the preliminary result with the highest PSNR and SSIM in four preliminary results.

Apparently we can find that the PCFC-Net greatly improves the quality of the inpainting results, under different proportions and sizes of the holes, the PSNR and SSIM are both improved by a certain degree, especially the PSNR, which indicates that PCFC-Net plays a crucial role in the proposed method and is extremely necessary to improve the video coherence and scene restoration of video inpainting.

Table 1. PSNR/SSIM under different proportions and sizes of holes.

Type	Proportion	10% (1/4)	25% (1/4)	10% (1/8)	25% (1/8)
PSNR	Coarsely stage	29.9	29.4	29.8	29.2
	PCFC-Net	34.8	33.9	36.0	34.6
SSIM	Coarsely stage	0.18	0.13	0.16	0.11
	PCFC-Net	0.73	0.66	0.80	0.72

Table 2. PSNR/SSIM under different proportions of holes.

Type	Proportion	5%	10%	15%	20%	25%	30%
PSNR	Kim et al.	33.3	33.2	33.1	32.9	32.7	32.7
	Ours	36.1	35.6	35.2	34.8	34.4	34.0
SSIM	Kim et al	0.71	0.71	0.69	0.67	0.66	0.65
	Ours	0.80	0.78	0.76	0.73	0.70	0.68

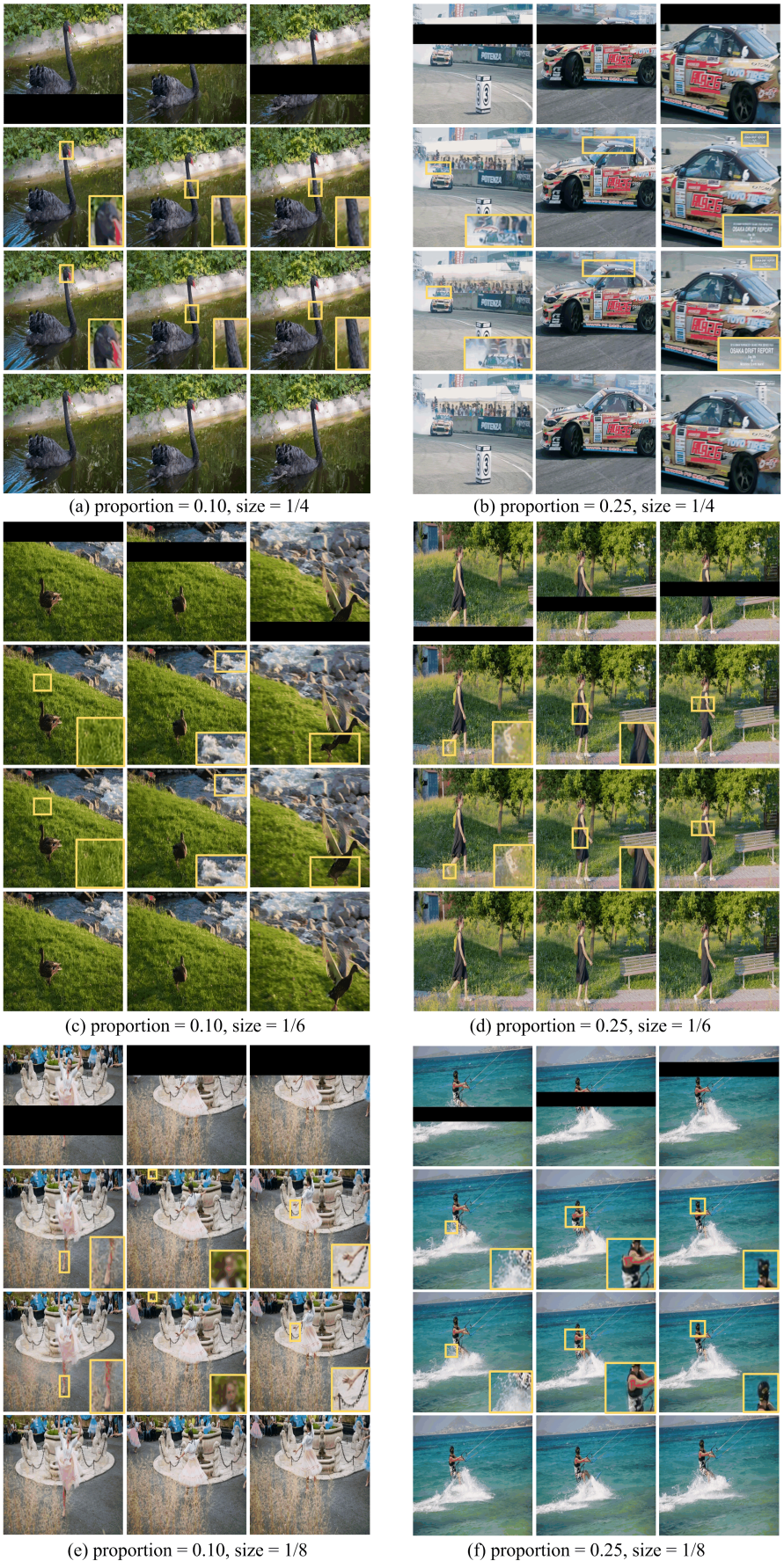
As shown in Table 2, when the size of holes is 1/6 of the frames, the PSNR score of our method is higher than the baseline under different proportion of holes, demonstrating that the inpainting results of our method have higher quality than the baseline. And so as the SSIM score, which illustrates that the inpainting results of our method are more similar to the ground-truth. Furthermore, with the proportion of holes increases, the PSNR and SSIM of the inpainting results of our method respectively decrease from 8.4% to 4.0% and from 12.7% to 4.6%, which is mainly caused by the high probability of the damage of reference frames P_{t+1}, P_{t+2}, but when there are holes nearly on every frame (proportion=30%), our method is still able to produce higher quality results than the baseline, which shows that our model is more suitable for the poor transmission condition.

We also compare baselines with different hole sizes to test the inpainting performance of our method when the transmitted video is damaged under different transmission strategies, as shown in Table 3, the PSNR is improved by 4.3% ~ 8.1%, and the SSIM is improved by 6.5% ~12.7%. Furthermore, we can also find that the smaller the size of holes, the more obvious the superiority of our method, which suggests us to divide the frames into small pieces when using the proposed method in unstable transmission for better visual effects.

Table 3. PSNR/SSIM under different sizes of holes.

Type	Proportion	10% (1/4)	25% (1/4)	10% (1/8)	25% (1/8)
PSNR	Kim et al	32.9	32.5	33.3	32.8
	Ours	34.8	33.9	36.0	34.6
SSIM	Kim et al	0.67	0.62	0.71	0.67
	Ours	0.73	0.66	0.80	0.72

The examples of inpainting results are shown in Figure 6, it is obvious that under different experimental settings, the visual effects of the inpainting results of our method are better than the baseline, which is reflected in four aspects: ①The inpainting results of our method are more consistent with the ground-truth; ②In the inpainting results of our method, the integrity of the objects in the inpainted frames is stronger, and the similarity and continuity of adjacent frames are also superior. ③ Our method performs better on the inpainting of details, which makes the inpainting results look sharper. ④ It should be noticed that even the content of the whole frame is damaged, the inpainting result of our method also achieves a satisfying performance.



(a) proportion = 0.10, size = 1/4 (b) proportion = 0.25, size = 1/4

(c) proportion = 0.10, size = 1/6 (d) proportion = 0.25, size = 1/6

(e) proportion = 0.10, size = 1/8 (f) proportion = 0.25, size = 1/8

Figure 6. Inpainting results under different proportion and size of holes. Row 1: The damaged frames; Row 2: The inpainting results of the baseline; Row 3: The inpainting results of our method; Row 4: The ground-truth.

Moreover, our method only requires two reference frames after the damaged frame, which is much shorter than the baseline (requiring the 6th frame after the damaged frame). If it takes T seconds to transmit one video frame, the baseline will wait $6T$ seconds to get the inpainting process started, while our method only needs $2T$ seconds, saving much time, especially in the case of unstable transmission where the worse the transmitting conditions, the longer it takes to transmit one frame. Therefore, it is obvious that our method is more suitable for inpainting the videos damaged in an unstable transmission than the baseline.

5 Conclusions

We propose a low-latency inpainting method against the completely random characteristics for video damage in unstable transmission, that can minimize the inpainting latency while ensuring the high quality of the inpainting result. Firstly, we start by using a small sampling window to select the reference frames, and use the masked FlowNet 2.0 to extract the reference optical flows, then based on the temporal continuity of motions, the estimate flows are calculated which is then used to warp the reference frames to the preliminary inpainting results of the damaged frame. Secondly, we propose the PCFC-Net which is designed as an encoder-decoder structure to make use of the temporal and spatial information of the reference frames, estimated flows and the preliminary inpainting results in order to generate the final inpainting result of the damaged frame. Validated on the DAVIS dataset, the proposed model consistently outperforms the state of the art baseline under different sizes and proportions of holes, under the condition that the length of the reference frame sampling window is reduced significantly.

Acknowledgments

This work is supported by the National Key Research and Development Project (Grant No. 2018AAA0100802), the National Natural Science Foundation of China (Grant No. 61374073), and the Anhui Provincial Natural Science Foundation (Grant No. 2008085MF198).

Conflict of interest

The authors declare no conflict of interest.

Author information

Wei Yutong is currently a master candidate in the Department of Automation under the supervision of Prof. Zhu Jin at University of Science and Technology of China. Her research mainly focuses on deep learning.

Bao Bingkun is currently a professor at the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications. Her research mainly focuses on multimedia computing, computer vision, etc.

Zhang Ziqi currently works at Agora in Shanghai.

Zhu Jin is currently an associate professor at the Department of Automation, University of Science and Technology of China. His research mainly focuses on filtering and control theory of stochastic systems.

References

[1] CISCO. Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.pdf. 2019.

[2] Alatas O, Yan P, Shah M. Spatio-temporal regularity flow (SPREF): Its estimation and applications. IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(5): 584-589.

[3] Shih T K, Tang N C, Hwang J N. Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(3): 347-360.

[4] Chung B, Yim C. Bi-sequential video error concealment method using adaptive homography-based registration. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(6): 1535-1549.

[5] Wang C, Huang H, Han X, et al. Video inpainting by jointly learning temporal structure and spatial details. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: IEEE, 2019, 33: 5232-5239.

[6] Xu R, Li X, Zhou B, et al. Deep flow-guided video inpainting. 2019, arXiv:1905.02884.

[7] Kim D, Woo S, Lee J Y, et al. Deep video inpainting. Proceedings of the Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 5792-5801.

[8] Ilg E, Mayer N, Saikia T, et al. FlowNet 2.0: Evolution of optical flow estimation with deep networks. Proceedings of the Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 2462-2470.

[9] Johnson J, Alahi A, Li F F. Perceptual losses for real-time style transfer and super-resolution. European Conference on Computer Vision. Amsterdam, Netherlands: IEEE, 2016: 694-711.

[10] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 1-9.

[11] Xu N, Yang L, Fan Y, et al. Youtube-VOS: Sequence-to-sequence video object segmentation. Proceedings of the European Conference on Computer Vision. Munich, Germnay: IEEE, 2018: 585-601.

[12] Pont-Tuset J, Perazzi F, Caelles S, et al. The 2017 DAVIS challenge on video object segmentation. 2017, arXiv:1704.00675.

(Continued on p. 786)

Research progress of interfacial mechanical behavior and design of nanocellulose-based sequentially architected materials

SONG Rongzhuang[1,2], HOU Yuanzhen[1,2], HE Zezhou[1,2],
XIA Jun[1,2], ZHU Yinbo[1,2*], WU Hengan[1,2]

1. Department of Modern Mechanics, University of Science and Technology of China, Hefei 230027, China;
2. CAS Key Laboratory of Mechanical Behavior and Design of Materials, University of Science and Technology of China, Hefei 230027, China
* Corresponding author. E-mail: zhuyinbo@ustc.edu.cn

Abstract: Nanocellulose exhibits superior mechanical properties and is a renewable natural biomass material. Nanocellulose-based sequentially architected materials are expected to become a new generation of environment-friendly high-performance structural and functional materials leading sustainable development. The construction of reasonable multiscale nonlinear coupling relationship between interfacial mechanical behavior and material microstructure is pivotal to the strengthening-toughening design of nanocellulose-based materials. Recent research progress of interfacial mechanical behavior and design of nanocellulose-based sequentially architected materials was reviewed here. The interfacial hydrogen-bonding behavior, multiscale interfacial mechanics, and some design cases of interfaces and microstructures were discussed. At last, the summary and perspective of key points in this field were given. This paper would aim to provide new perspectives for the design and preparation of high-performance nanocellulose-based sequentially architected materials based on micro-nano mechanics and multiscale mechanics.

Keywords: nanocellulose; multiscale mechanics; hydrogen bonds; interface; ordered microstructure; material design

(Continued from p. 724)

不稳定传输中受损视频的低延迟修复方法

魏俣童[1],鲍秉坤[2],张子祺[3],朱进[1*]

1. 中国科学技术大学自动化系,安徽合肥 230027;
2. 南京邮电大学通信与信息工程学院,江苏南京 210003;
3. 声网公司,上海 200082
* 通讯作者. E-mail:jinzhu@ustc.edu.cn

摘要: 视频流量已逐渐成为移动流量的重要组成部分,而不稳定传输中的视频缺损却仍然是一个亟待解决的问题. 这种类型的视频缺损往往带有完全随机的特性,很难对其进行低延迟并且高精度的修复. 我们率先关注了该不稳定传输中视频修复的任务,并提出了一种低延迟的视频修复方法,该方法包括两个阶段:在粗略修复阶段,先从参考帧中提取受损的二维光流图,再建立线性预测模型,根据运动在时间维度的连续性,来对受损帧进行初步的粗略修复. 在精细修复阶段,提出了一个部分卷积神经网络(PCFC-Net),用于对所有参考信息进行综合并计算精细修复的结果. 与基线相比,该方法在DAVIS数据集上的参考帧等待时间大大减少,同时PSNR和SSIM也提高了4.0%~12.7%.

关键词: 视频修复;不稳定传输;部分卷积神经网络;线性预测