SCIENCE CHINA
Information Sciences

• LETTER •

July 2022, Vol. 65 179103:1–179103:2
https://doi.org/10.1007/s11432-019-2956-6

3DPF-FBN: video inpainting by jointly 3D-patch filling and neural network refinement

Yan HUANG, Chuanchuan YANG* & Zhangyuan CHEN

State Key Laboratory of Advanced Optical Communication Systems and Networks, Department of Electronics, Peking University, Beijing 100871, *China*

Received 23 November 2019/Revised 5 March 2020/Accepted 19 June 2020/Published online 14 May 2021

Citation Huang Y, Yang C C, Chen Z Y. 3DPF-FBN: video inpainting by jointly 3D-patch filling and neural network refinement. Sci China Inf Sci, 2022, 65(7): 179103, https://doi.org/10.1007/s11432-019-2956-6

Dear editor,

Video inpainting is a new technology of recovering the lost motion vectors or image blocks during the video sequence transmission. It aims to speculate the correct value of missing voxels from a given video to achieve the automatic inpainting and make the inpainted video frames present a natural visual effect. However, video inpainting remains challenging owing to the difficulties coming from computational complexity, complex scenarios and the high requirement on spatial and temporal consistency.

Video inpainting requires not only recovering the missing areas within each frame but also maintaining the content consistency between successive frames. Existing video inpainting can generally fall into three categories: patch-based method [1], object-based method [2] and deep-learning based method [3]. The patch-based method formulates the inpainting problem as image inpainting, which uses appropriate pixels to fill in the missing region through sampling spatial patches from neighborhoods of the missed area. Object-based method requires pre-processing first to automatically segment the input video into static background and moving foreground, and then get down to repair their missing parts individually and finally synthesize the two reconstructed results into an inpainted video. Deep-learning based method tends to develop an efficient neural network for learning image feature to predict the contents in the occlusions.

Object-based methods usually require specific situations like that periodic motion. The synthesized step of moving objects and static background may cause temporal inconsistency. As a result, most inpainting studies are likely to focus on patch-based or learning-based methods. Typical patch-based methods depend on spatial pixels and patch matching, and they are weak in temporal heterogeneous cases while learning-based models can handle temporal consistency in image prediction and inference. Despite some better achievements, learning-based methods may limit their practicality and flexibility in complex scenarios such as large span motion or dynamic motion from multiple objects. Therefore, owing to these challenging issues, exploring feasible combination on different inpainting types shows a great potential for development and improvement in terms of the inpainting accuracy on various challenging situations.

Framework. To get competitive results in various challenging cases such as moving background, complex motion or long-lasting occlusion, a method called 3DPF-FBN by jointly 3D-patch filling and forward-back network refinement is utilized in this study. This framework consists of two parts. (1) 3DPF: searching for appropriate 3D patches from neighbor content and incorporating them into still-unknown regions in the video. (2) The inpainted area in the output of (1) will be refined and optimized to the fullest with a forward-back network (FBN) to achieve a thorough inpainted result. The motivation behind the method is that patch-based inpainting can complete a rapid filling with suitable 3D patch directly, and deep-learning network can facilely propagate visually realistic pixels to generate inference frames that preserve the temporal coherence naturally.

3DPF. In this study, the patch-based method in Le et al. [4] is further improved to pursue a preferable efficiency with two important innovations. We define the input video sequence as a 3D spatial-temporal cubical space and change the optical flow to the 3D preprocessing step and propagation step. A foreground/background patch-cubic clustering is utilized to achieve the global search in the whole cubical space, which makes the search stride adjustable. For the 3D patch matching, we argue not to implement the random searching for every occluded pixel, but pay attention to that pixels on a sparse grid without abandon its efficiency.

FBN. Optical flow plays a great role in video-related tasks as it can provide direct access to the temporal information in successive frames. FBN aims to generate robust flow-features from the input video as an LSTM sequence into the ConvLSTM architecture and interpolation computing to keep the output temporal motion coherent in frame prediction. It can achieve a forward frame inference and a backward motion update for modeling the spatial-temporal consistency in videos, which is applied to fulfill a coarse-to-fine refinement on the inpainted results from 3DPF.

* Corresponding author (email: yangchuanchuan@pku.edu.cn)

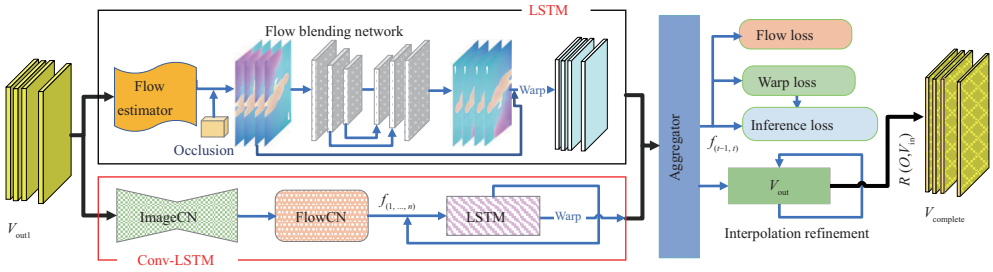

Figure 1 (Color online) Procedure of the proposed FBN inpainting refinement.

This study is built upon a global patch-based inpainting and deep-neural network refinement. In the 3DPF stage, it takes an incomplete video V_{in} and a mask occlusion M as input, and produces an inpainted video V_{out1} as output, which will be input into the FBN stage to complete the refinement. It takes a total of 5 frames as input at each iteration, 4 source frames and 1 reference frame (i.e., the reference frame is the frame inpainted with 3DPF and it is also the frame that will be refined with FBN), and generates a forward refinement frame similar to the reference frame. A backward interpolation frame among the two frames is updated to be the inpainted frame, in which the under-inpainted regions will be used to fill in the occlusions of V_{in} to complete the refinement.

To better exploit temporal and motion context coming from sequential frames, we integrate the modification of optical flow into the 3D-2D encoder-decoder model FlowNet [5], which can explicitly reveal traceable features from the video dynamics. The network learns each flow feature between consecutive frames to deal with both pixel propagation and hole-filling while preserving the motion coherence. Figure 1 illustrates the FBN refinement structure. The flow feature is extracted from V_{out1} with FlowNet [5], which will be mapped into the Flow Blending Network [6], where the extracted features from the encoder are concatenated to the corresponding decoder layers to enhance the performance and then facilitate the final evaluation of the inference loss.

Training losses and strategy. In the proposed neural network, the overall loss function of training the model is defined as

$$L_{\text{total}} = \alpha L_r + \beta L_p + \gamma L_f, \tag{1}$$

where L_r denotes the reconstruction loss; i.e., we compute the L1 distance to guide our model spatially reconstructing the frames. The L_p loss focuses on the pixel-level features. L_f is the flow estimation loss. α, β, γ denote the balancing weights for L_r, L_p, L_f loss, respectively, which are manually pre-defined to 1, 2, 10 respectively throughout the experiments. Specially, the L_r and L_p loss are

$$L_r = \sum_{t=2}^{n} M_t \cdot \left\| O_{t,x,y} - O'_{t,x,y} \right\|_1, \tag{2}$$

$$L_p = \sum_{t=2}^{n} M_t \cdot \left\| O_{t,x,y} - V'_{t,x,y} \right\|_1, \tag{3}$$

where n, x, y are the spatial temporal coordinates of the video. The flow loss L_f is defined as

$$L_f = \sum_{t=2}^{n} \left\| V_{t(L\to 1)} - V'_{t(L\to 1)} \right\| + \left\| F_t - V'_{t(L\to 1)} F_{t-1} \right\|_1, \tag{4}$$

in which $V_{t(L\to 1)}$ is the flow feature of level L to 1 in the target frames, f_t and f_{t-1} are processed in that flow blending network. In order to facilitate the optimization of inpainting, before training the flow-based network, we have pre-trained the flow inpainting and frames inpainting separately to ensure a stable status. Furthermore, we augment the training dataset by extending the flip horizontal and shift transformation of the frames for the sake of generalization ability in our training model. With all of the training losses and refinement strategy, our video flow-based network is able to optimize frames with plausible details.

Conclusion. We propose a novel video inpainting approach called 3DPF-FBN, which is the first to incorporate patch-based filling and learning-based network for the general video inpainting task. It can largely facilitate inpainting videos in many conditions such as arbitrary missing holes, complex motions, and yet maintain spatial and temporal consistency. Our experimental results demonstrate its effectiveness on both the DAVIS and YouTube-VOS datasets with competitive performance, and it presents great potential on both inpainted results and total computation time.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2018YFB1801702) and Joint Fund of the Ministry of Education (Grant No. 6141A02033347).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Koloda J, Seiler J, Peinado A M, et al. Scalable kernel-based minimum mean square error estimator for accelerated image error concealment. IEEE Trans Broadcast, 2017, 63: 59–70
2 Granados M, Tompkin J, Kim K, et al. How not to be seen — object removal from videos of crowded scenes. Comput Graph Forum, 2012, 31: 219–228
3 Liu G, Reda F A, Shih K J, et al. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision, Munich, 2018. 85–100
4 Le T T, Almansa A, Gousseau Y, et al. Motion-consistent video inpainting. In: Proceedings of the IEEE International Conference on Image Processing, Beijing, 2017. 2094–2098
5 Ilg E, Mayer N, Saikia T, et al. Flownet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 2462–2470
6 Ding Y, Wang C, Huang H, et al. Frame-recurrent video inpainting by robust optical flow inference. 2019. ArXiv: 1905.02882