Temporal Adaptive Alignment Network for Deep Video Inpainting

Ruixin Liu, **Zhenyu Weng**, **Yuesheng Zhu*** and **Bairong Li**
Communication and Information Security Lab, Shenzhen Graduate School, Peking University
{anne_xin, wzytumbler, zhuys, lbairong}@pku.edu.cn

Abstract

Video inpainting aims to synthesize visually pleasant and temporally consistent content in missing regions of video. Due to a variety of motions across different frames, it is highly challenging to utilize effective temporal information to recover videos. Existing deep learning based methods usually estimate optical flow to align frames and thereby exploit useful information between frames. However, these methods tend to generate artifacts once the estimated optical flow is inaccurate. To alleviate above problem, we propose a novel end-to-end Temporal Adaptive Alignment Network(TAAN) for video inpainting. The TAAN aligns reference frames with target frame via implicit motion estimation at a feature level and then reconstruct target frame by taking the aggregated aligned reference frame features as input. In the proposed network, a Temporal Adaptive Alignment (TAA) module based on deformable convolutions is designed to perform temporal alignment in a local, dense and adaptive manner. Both quantitative and qualitative evaluation results show that our method significantly outperforms existing deep learning based methods.

1 Introduction

Video inpainting is a task of synthesizing visually realistic and semantically plausible contents in missing regions of the given video sequence in temporal coherence. It can be used in many applications such as unwanted object removal, damaged parts recovery, visual privacy filtering, etc. Significant progress has been made recently in single-image inpainting[Pathak *et al.*, 2016; Yu *et al.*, 2018] thanks to the deep generative networks. However, because of the additional time dimension, video inpainting method not only needs to repair the missing region for each frame but also has to ensure the temporal consistency across frames. While temporal relationship brings challenges for video inpainting, the temporal redundancy information can be exploited by it to obtain better results at the same time. As illustrated in Figure 1, compared with image inpainting based method, video inpainting method performs better results and preserves the video coherence. However, it is difficult to directly utilize information from reference frames due to the misalignment between frames which is caused by complex motion of camera or objects. Therefore, how to efficiently utilized temporal information is the essential issue for video inpainting problem.

Figure 1: (a) Input video with masks in red. (b) Video inpainting results from image inpainting method [Nazeri *et al.*, 2019]. (c) Our video inpainting method.

Traditional patch-based methods[Newson *et al.*, 2014; Huang *et al.*, 2016] find the similar spatio-temporal patches from the known regions of videos to fill the holes, which formulate the problem as a patch-based optimization task. Although some good results have been shown, they usually suffer from the high computational complexity, which results in very slow processing speed.

Motivated by the success of the deep neural networks in single image inpainting task, several deep learning based methods for video inpainting have been proposed recently and achieve significant results in terms of quality and speed. [Lee *et al.*, 2019] align frames firstly by performing global affine transformation, and copy valuable pixels from aligned frames to complete the missing regions. It shows that tem-

*Contact Author

poral alignment is important for exploiting information from reference frames. However, such transformation could not well align the frames with more complex motions, which are common in realistic scenes.

Optical flow contributes to obtaining temporal information between frames in complex motion scenes. [Xu *et al.*, 2019; Kim *et al.*, 2019] utilize the optical flow to align frames, visible information from reference frames is collected via flow warping operation. These methods highly depend on the accuracy of the estimated optical flow. However, completing optical flow between frames with missing region is quite challenging and any errors in optical flow computation will influence the final quality of results.

In this paper, a novel end-to-end Temporal Adaptive Alignment Network(TAAN) without using optical flow is suggested by us to solve the video inpainting problem. Given a sequence of video frames with holes, the network processes video frame-by-frame. The TAAN aligns target frame with reference frames firstly at a feature level without explicit motion estimation through a TAA module. Specifically, inspired by the deformable convolution [Dai *et al.*, 2017], the proposed TAA module utilizes the features from target frame and corresponding reference frame to predict the offsets of sampling convolution kernels, and applies the kernels on the reference frame features to perform temporal adaptive alignment. In this way, the final reconstructed target frame will have less artifacts and the capability of handling various motion conditions in temporal scenes will be improved. In addition, benefited by deformable convolution, more information with the same structure as the sampled position will be explored by TAA module, which strengthens the alignment accuracy between frames. Finally, a reconstruction network which takes the aggregated aligned reference frame features and target features is developed to recover target frame. The end-to-end network design helps our network pay more attention to the missing region and further improve the performance of the model.

We conduct extensive experiments on Youtube-VOS [Xu *et al.*, 2018] and DAVIS [Perazzi *et al.*, 2016] datasets. The experimental results show that our framework could achieve visually pleasing results at complex motion scenes. The major contributions of our paper are summarized as follows:

- We propose a novel end-to-end network for video inpainting, which aligns the frames at a feature level via implicit motion estimation and aggregates temporal features to synthesize missing content.
- We propose a temporal adaptive alignment module based on standard deformable convolution to adaptively align frames which contain holes in a local and dense manner.
- We quantitatively and qualitatively evaluate our method and show its efficacy.

2 Related Work

2.1 Traditional Methods

Traditional patch-based approaches typically use the prior such as patch similarity to propagate information from the known regions. [Patwardhan *et al.*, 2005; Patwardhan *et al.*, 2007] complete videos by sampling non-local spatio-temporal patch assuming static camera and constrained camera motion. They find the patch based on the greedy algorithm, which inevitably propagate the early errors and lead to globally inconsistent results. To enforce global spatio-temporal consistency, [Wexler *et al.*, 2007] cast the problem as global optimization task which constrain missing values form coherent structures with respect to reference examples. Further, to strengthen the temporal consistency, [Newson *et al.*, 2014] use a robust affine estimation to compensate camera motion and perform an extension of PatchMatch algorithm [Barnes *et al.*, 2009] to accelerate the patch matching process. [Granados *et al.*, 2012] compensate geometric distortion by utilizing a set of homographies to perform frame-to-frame alignments. However, they have difficulties in handling general scenes which have more complex geometric variations. [Huang *et al.*, 2016] combine the advantages of optical flow and color information, and formulate the problem as a global optimization of color and flow. They perform significantly results in complex motion scenes, achieving the state-of-the-art results.

2.2 Learning-based Methods

A significant advantage of deep learning based methods is their ability to learn semantics from large scale datasets. The first deep learning based method for video inpainting is proposed by [Wang *et al.*, 2019], they perform 3D convolution on low resolution input to provide the temporal guidance and use 2D convolution based on the low resolution result to recover the spatial coherent frame. [Chang *et al.*, 2019] propose 3D gated convolution network for video inpainting based on the work of image inpainting, which tackle the uncertainty of free-form masks problem. Further, they introduce a Temporal PatchGAN discriminator to enhance temporal consistency. However, 3D convolution is hard to train and their temporal receptive fields are too limited or directional. [Kim *et al.*, 2019] collect information by flow warping from neighbor frames to the target frame to recover the target frame, and utilize a recurrent feedback and memory layer to stable temporal consistency. [Xu *et al.*, 2019] address video inpainting problem as a pixel propagation problem by recovering accurate flow field from missing region and the synthesized flow field is used to guide the pixel propagation to generate semantically plausible contents. Despite the fact that motion can be the useful guidance for propagating information [Oh *et al.*, 2019] aggregate temporal information through an asymmetric attention block to progressively fill the hole from the hole boundary. [Lee *et al.*, 2019] utilize a self-supervised network to estimate affine matrices between frames for the alignment and aggregate information from aligned frames through a copy-and-paste network to fill the missing regions.

3 Proposed Algorithm

3.1 Overview

Figure 2 presents the workflow of the proposed video inpainting network. Given a sequence of video frames X annotated with missing region M(1 indicates invalid pixels, 0 on

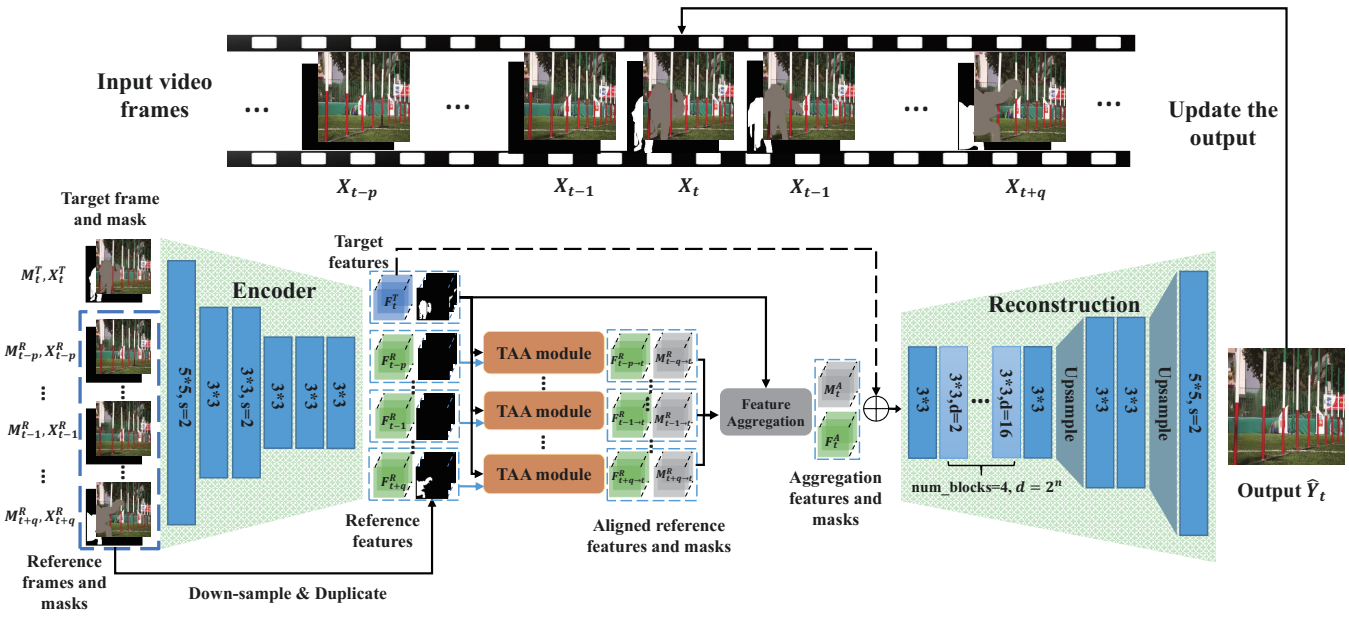

Figure 2: The schematic illustration of the our TAAN-based video inpainting model.

the contrary), the network processes video frame-by-frame in the temporal order and generates the final outputs $\widehat{Y}$. The whole network consists of four modules: encoder module, TAA module, feature aggregation module and reconstruction module.

To complete a target frame, our network first takes the target frame X^T, reference frames X^R and corresponding binary masks into an encoder module to extract feature maps F. All of them share the same encoder. [Lee *et al.*, 2019] have demonstrated that taking account into the missing region when aggregating information will help improve the effectiveness of inpainting results. So we down-sample the mask M and then extend it to the same number of channels as the extracted feature maps by duplication operation. Next, we send the feature maps that extracted from the target frame X_t^T and the reference frame X_i^R, corresponding duplicated masks MD into TAA module to get aligned results ($F_{i\to t}^R$ and $M_{i\to t}^R$, where $i \to t$ indicates reference features/masks i is aligned to target t). After alignment, all aligned reference frame features and aligned masks are aggregated by feature aggregation module. Finally, we concatenate target features F_t^T, aggregated features F^A and aggregated masks M^A to recover the target frame.

To enforce the temporal consistency, we update the input video sequence with completed frame over time and transfer the previous completed frame as one of the reference frames of the current target frame.

3.2 Network Architecture

Encoder

This module concatenates the frame and corresponding binary mask along the channel axis to form a 4-channel image and take it as input to extract visual feature maps. The convolutions with stride of 2 are utilized to decrease the resolution twice and get the 1/4 scale of original size, which is important to maintain the high frequency details in the missing region [Iizuka *et al.*, 2017]. The extracted features will be employed for feature-wise temporal alignment.

Temporal Adaptive Alignment(TAA) Module

By learning offsets of the sampling convolution kernels and applying the learned kernels to the feature maps, deformable convolution [Dai *et al.*, 2017] could obtain information away from its regular local neighborhood, improving the complex geometric transformation capability. Motivated by the capability of the deformable convolution, we introduce it into our TAA module to perform temporal alignment. A detailed illustration of our TAA module is presented in Figure 3.

Given the visual features F_t^T and F_i^R from target frame and reference frame respectively, TAA module concatenates these features and feeds them into an offset estimator network to predict offsets $\{\Delta p_n \mid n = 1, \cdots, |\mathcal{R}|\}$ of the convolution kernels. For example, $\mathcal{R} = \{(-1,-1),(-1,0),\cdots,(0,1),(1,1)\}$, when a regular grid with a 3×3 kernel, and a dilation factor of 1. The existence of missing regions bring challenges for alignment. To solve this problem, we adopt the U-Net architecture for the offset estimator network which has been widely used in pixel-wise estimation tasks. Further, a 3×3 convolution layer is utilized to predict the final offsets of the reference features and duplicated mask.

With the predicted offsets $\{\Delta p_n\}$ and reference frame features F_i^R, each position p_0 in the aligned feature maps $F_{i\to t}^R$ are computed by the deformable convolution operation as follows:

$$y(p_o) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n + \Delta p_n). \qquad (1)$$

It is worth noting that the adaptively learned offsets will

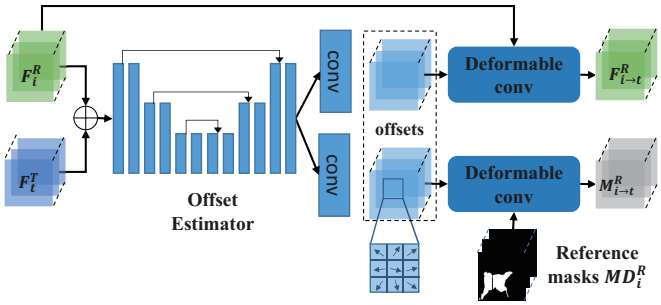

Figure 3: The schematic illustration of how the TAA module aligns features and masks.

implicitly capture motion information and contribute to aligning frames in the complex motion scenes. In addition, as illustrated in Figure 3, for each sampled position p_0, our TAA module could explore more features that may share the same image structure as p_0 by deformable convolution operation, which helps to collect more information in reference frame features and further improve alignment capability.

The aligned masks are computed by the offsets and duplicated masks in a similar fashion. For better results, we use the DCNv2 [Zhu *et al.*, 2019] in our implements, which has more stronger modeling power. We will analyze the effectiveness of TAA module in Sec.4.3.

Feature Aggregation

Different frames and locations are not equally beneficial to the reconstruction. Inspired by strong results presented in [Lee *et al.*, 2019], we pick up the most relevant information in the reference frames when aggregating the features as follows:

We first measure the global similarities $\theta_{i,t}$ between the target frame feature and each aligned reference frame feature, while ignoring the invalid location in the aligned mask.

$$\theta_{i,t} = \frac{1}{\sum_{(x,y)} \boldsymbol{V}_{i,t}(x,y)} \cdot \sum_{(x,y)} \boldsymbol{V}_{i,t}(x,y) \cdot \boldsymbol{F}_t^T(x,y) \cdot \boldsymbol{F}_i^R(x,y). \tag{2}$$

Where $\boldsymbol{V}_{i,t} = (1 - \boldsymbol{M}\boldsymbol{D}_t^T) \odot (1 - \boldsymbol{M}_{i\to t}^R)$ is the visibility map.

Then, we multiply the similarity with corresponding aligned masks and use a softmax function across temporal dimension to weigh the features in the aligned reference frames. The weight is computed as follows:

$$\boldsymbol{W}_i(x,y) = \frac{exp(\theta_{i,t} \cdot (1 - \boldsymbol{M}_{i\to t}^R))}{\sum_i exp(\theta_{i,t} \cdot (1 - \boldsymbol{M}_{i\to t}^R))}. \tag{3}$$

The final fusing features are computed by summing the reference frames features with the weight.

$$\boldsymbol{F}^A(x,y) = \sum_i \boldsymbol{F}_{i\to t}^R \cdot \boldsymbol{W}_i(x,y). \tag{4}$$

The aggregation masks are computed as follows:

$$\boldsymbol{M}^A(x,y) = 1 - (\sum_i \boldsymbol{W}_i(x,y)). \tag{5}$$

Reconstruction

The reconstruction network is used to restore the target frame by taking the aggregated reference features, aggregated reference masks and the target features as input. Four dilation blocks with dilation factor of 2^n are designed to enlarge the receptive field, which is beneficial to fill the region not existing in the reference frames [Iizuka *et al.*, 2017]. Finally, the nearest neighbor up-sampling is used to enlarge the feature map to the target frame.

3.3 Loss Functions

To guarantee the completion quality of the results, we adopt reconstruction loss, perceptual loss, style loss and total variation loss function to effectively train the proposed network.

Reconstruction loss is used to constrain pixel-level restoration. Where I_{gt} represents the ground truth image, M is given mask, I_{out} is the network prediction.

$$\mathcal{L}_{hole} = \|M \cdot (I_{out} - I_{gt})\|_1 . \tag{6}$$

$$\mathcal{L}_{valid} = \|(1 - M) \cdot (I_{out} - I_{gt})\|_1 . \tag{7}$$

We also adopt perceptual loss $\mathcal{L}_{prec}$ and style loss $\mathcal{L}_{style}$ to further improve the visual quality of the whole recovered images.

$$\mathcal{L}_{prec} = E\left[\sum_i \frac{1}{N_i} \|\phi_i(I_{gt}) - \phi_i(I_{out})\|_1\right]. \tag{8}$$

$$\mathcal{L}_{style} = Ej\left[\left\|G_j^\phi(I_{out}) - G_j^\phi(I_{gt})\right\|_1\right]. \tag{9}$$

Where ϕ_i is the activation map in a pretrained VGG-19 on ImageNet. In our paper, ϕ_i corresponds to activation maps from layers relu1_1, relu2_1, relu3_1, relu4_1 and relu5_1. The activation maps also will be sent to calculate gram matrix in style loss. G represents Gram Matrix for computing covariance.

Furthermore, we utilize the total variation loss to smooth the checkerboard effect. The total loss $\mathcal{L}_{total}$ is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{valid} + 6\mathcal{L}_{hole} + 0.05\mathcal{L}_{prec} + 120\mathcal{L}_{style} + 0.1\mathcal{L}_{tv}. \tag{10}$$

For our experiments, the loss term weights are adopted from [Liu *et al.*, 2018].

4 Experiments

4.1 Experimental Settings

Datasets. Keeping with the goal of synthesizing plausible contents of the missing region in videos, two datasets are employed in this work to demonstrate the effectiveness of the proposed method. The first is YouTube-VOS [Xu *et al.*, 2018] Dataset, which is a large-scale video object segmentation dataset with a wide variety of scenes. It contains 4,453 YouTube video clips and 94 object categories and is split into 5,471 for training, 474 for validation and 508 for testing. We train our video inpainting network on the YouTube-VOS training set. The second dataset is DAVIS [Perazzi *et al.*, 2016] Dataset. It includes 50 high quality video sequences with covering dynamic scenes, large occlusions, motion blur, complex camera movements, and each frames are annotated with the pixel-accurate foreground object masks.

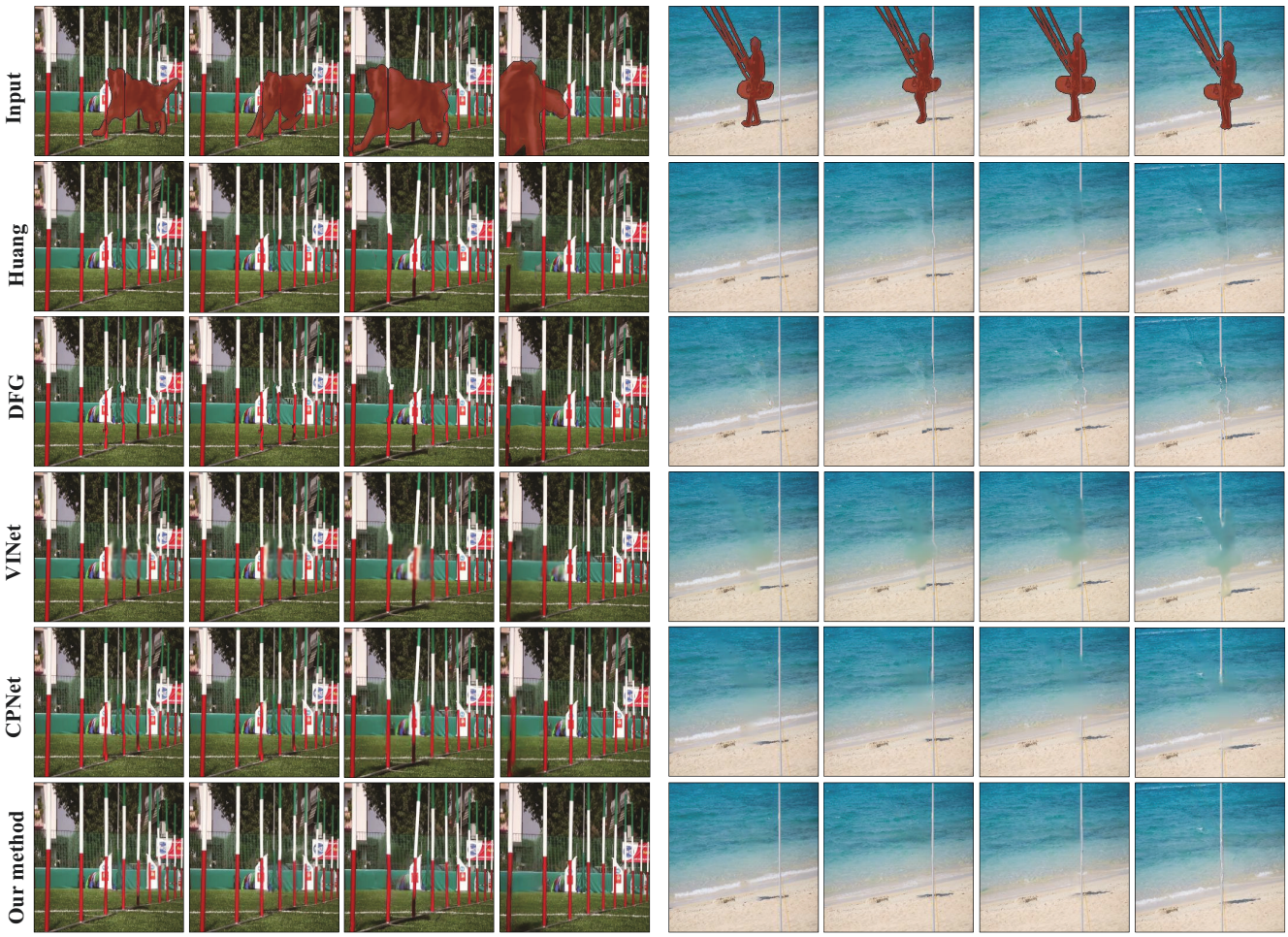

Figure 4: Qualitative comparison results of TAAN video inpainting approach for the scenes dog-agility(left) and kite-walk(right) from DAVIS Dataset.

Training Details. We select five reference frames(X_{t-4}, X_{t-2}, X_{t-1}, X_{t+2}, X_{t+4}) and resize them into 256×256 as inputs when training the network. To accelerate the training process while reducing over-fitting, we initialize parameters of our neural network by using the initialization method in [He *et al.*, 2015]. Adam optimizer with the initial learning rate to 10^{-4} is utilized, we decayed the learning rate by 0.1 every 1 million iterations.

Baseline. We compare the proposed algorithm with the state-of-the-art methods including a traditional patch-based method:Huang [Huang *et al.*, 2016] and three deep learning based methods:DFG [Xu *et al.*, 2019], VINet [Kim *et al.*, 2019], CPNet [Lee *et al.*, 2019]. For deep learning based methods, we directly conduct the experiment on the trained model provided by authors. All experiments are done on the 256×256 frames.

4.2 Experimental Results

Qualitative Evaluations

To validate the generalization ability of our model, we compare the proposed method with other methods in real world scenes. The results as shown in Figure 4.

We can observe that the optical flow based methods [Xu *et al.*, 2019; Kim *et al.*, 2019] always tend to produce artifacts due to the wrong estimation of flow. CPNet adopts simple global motion estimation to align frames, which limits its performances for scenes with complex motion(e.g., scene dog-agility: the rods produce deformation over time).

Our model adaptively aligns frames in a local and dense manner at feature level, which can implicitly capture motion cues and aggregate more information to perform alignment, showing favorably inpainting results.

Quantitative Evaluations

Since there is no existing dataset to evaluate the video inpainting task. When testing on the Youtube-VOS dataset, we apply the video mask generation algorithm [Chang *et al.*, 2019] to simulate masks of the holes and synthesize videos by applying the generated masks on the background image sequences. Our method is supposed to recover the original videos. DAVIS dataset provides the dense object mask annotations. To get closer to the reality, we directly shuffle the

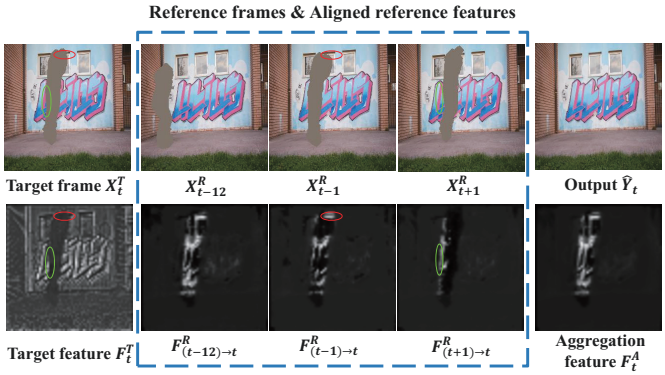

Figure 5: The effectiveness of TAA Module. The gray hole in the first row indicates the missing area.

Method	DAVIS		Youtube-VOS	
	PSNR	SSIM	PSNR	SSIM
Huang	30.607	**0.927**	29.015	0.885
DFG	29.557	0.911	27.333	0.879
VINet	29.765	0.901	27.542	0.874
CPNet	29.121	0.886	28.879	0.865
Ours	**30.829**	0.913	**29.987**	**0.887**

Table 1: Results of quantitative evaluation. The best result is labeled with **boldface**.

pairs of videos and masks from DAVIS to test the models when we conduct experiments on the DAVIS dataset.

The metrics results conducted on the Youtube-VOS and DAVIS datasets are summarized in Table 1. We can observe that our method is superior to all deep learning based methods. Although Huang [Huang *et al.*, 2016] achieves comparable results with our method, it formulates the video painting problem as a patch-based optimization task and is much slower than deep learning based methods, which has been indicated in our paper and other papers(e.g., [Xu *et al.*, 2019], [Kim *et al.*, 2019]).

4.3 Effectiveness of TAA Module

To validate the effectiveness of the proposed TAA module for alignment, the intermediate learned feature maps are visualized in Figure 5. As for target frame X_t^T, we selected 3 representative reference frames:X_{t-12}^R is far from the target frame, X_{t-1}^R and X_{t+1}^R are close to it.

The missing region of X_{t-1}^R and X_{t+1}^R overlap the target frame, so they only capture part of the information of the missing region in X_t^T. The green and red oval circles in target feature F_t^T are invisible, while $F_{(t-1)\to t}^R$ and $F_{(t+1)\to t}^R$ are able to predict the details, and our module will not capture corresponding information in the overlap region of target frame and reference frames. This shows that our module has excellent performance on frames alignment at the feature level. The aligned feature map $F_{(t-12)\to t}^R$ is similar to the aggregated feature map F_t^A, both of them obtain the fully details of the missing region, which indicates our TAA module has advantage on capture long temporal range of information

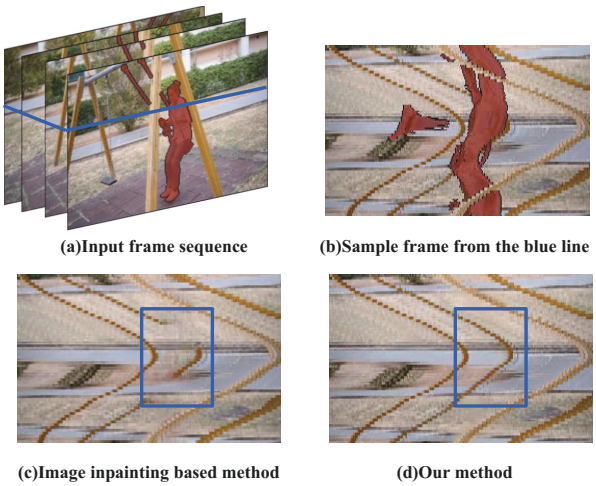

Figure 6: Temporal coherent completion.

and could effectively aggregate aligned feature map to help our network to reconstruction the frame. As shown in the figure, TAA module pays more attention to the missing region and further improve the performance of the model.

4.4 Temporal Consistency

To enforce temporal consistency, we update the input video frames with completed frame at each iteration, and propose TAA module to propagate information between consecutive frames via implicit alignment. We also process the frames one by one and run the network from the first frame to the last frame, then reverse the order during the test stage.

To show the temporal consistency between frames, we sample the slices of video frames along the blue line, and stack them along the vertical-axis as [Huang *et al.*, 2016]. In Figure 6, our method performs more smoother result than image inpainting based method [Nazeri *et al.*, 2019], which indicates our method shows better temporal consistency.

5 Conclusion

In this paper, we present an end-to-end temporal alignment network for video inpainting. A TAA module is proposed based on deformable convolution to perform temporal adaptive alignment in a feature domain without explicit motion estimation (e.g., optical flow). Our network significantly utilizes the temporal information from reference frames based on TAA module, which is important for video inpainting, and produces visually pleasing and temporally coherent results. In the future research, we will extend our framework to solve other video restoration tasks in practical applications.

Acknowledgements

This work was supported in part by NSFC-Shenzhen Robot Jointed Founding under Grant U1613215, in part by the Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing), and in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010137001.

References

[Barnes *et al.*, 2009] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24, 2009.

[Chang *et al.*, 2019] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9066–9075, 2019.

[Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017.

[Granados *et al.*, 2012] Miguel Granados, Kwang In Kim, James Tompkin, Jan Kautz, and Christian Theobalt. Background inpainting for videos with dynamic objects and a free-moving camera. In *European Conference on Computer Vision*, pages 682–695. Springer, 2012.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[Huang *et al.*, 2016] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (TOG)*, 35(6):196, 2016.

[Iizuka *et al.*, 2017] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017.

[Kim *et al.*, 2019] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019.

[Lee *et al.*, 2019] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4413–4421, 2019.

[Liu *et al.*, 2018] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision*, pages 85–100, 2018.

[Nazeri *et al.*, 2019] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[Newson *et al.*, 2014] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *SIAM Journal on Imaging Sciences*, 7(4):1993–2019, 2014.

[Oh *et al.*, 2019] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4403–4412, 2019.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[Patwardhan *et al.*, 2005] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmio. Video inpainting of occluding and occluded objects. In *IEEE International Conference on Image Processing*, pages 69–72, 2005.

[Patwardhan *et al.*, 2007] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmío. Video inpainting under constrained camera motion. *IEEE Transactions on Image Processing*, 16(2):545–553, 2007.

[Perazzi *et al.*, 2016] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

[Wang *et al.*, 2019] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5232–5239, 2019.

[Wexler *et al.*, 2007] Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time completion of video. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 29(3):463–476, 2007.

[Xu *et al.*, 2018] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[Xu *et al.*, 2019] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. pages 3723–3732, 2019.

[Yu *et al.*, 2018] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[Zhu *et al.*, 2019] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.