Align-and-Attend Network for Globally and Locally Coherent Video Inpainting

Sanghyun Woo[1]
shwoo93@kaist.ac.kr
Dahun Kim[1]
mcahny@kaist.ac.kr
Kwanyong Park[1]
pkyong7@kaist.ac.kr
Joon-Young Lee[2]
jolee@adobe.com
In So Kweon[1]
iskweon77@kaist.ac.kr

[1] Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
[2] Adobe Research, San Jose, CA, USA

Abstract

Video inpainting is more challenging than image inpainting because of the extra temporal dimension. It requires inpainted contents to be globally coherent in both space and time. A natural solution for this problem is aggregating features from other frames, and thus, existing state-of-the-art methods rely heavily on 3D convolution or optical flow. However, these methods emphasize more on the temporally nearby frames, and long-term temporal information is not sufficiently stressed. In this work, we propose a novel two-stage alignment method. The first stage is an alignment module that uses computed homography between the target frame and the reference frames. The visible patches are then aggregated based on the frame similarity to roughly fill in the target holes. The second stage is an attention module that matches the generated patches with known reference patches in a non-local manner to refine the previous global alignment stage. Both stages consist of large spatial-temporal window size for the reference and thus enable modeling long-range correlations between distant information and the hole regions. The proposed model can even handle challenging scenes with large or slowly moving holes, which have been hardly modeled by existing approaches. Experiments on video object removal demonstrate that our method significantly outperforms previous state-of-the-art learning approaches.

1 Introduction

Video inpainting aims to fill spatial-temporal holes with plausible content in a video. It is a practical and crucial problem as it could be beneficial for various video editing and restoration tasks. However, it is very challenging to maintain both spatial [1, 2, 3, 4, 5, 16, 22] and temporal consistency [6, 8, 12, 18, 23]; The inpainted contents must be spatially plausible, and temporally coherent at the same time.

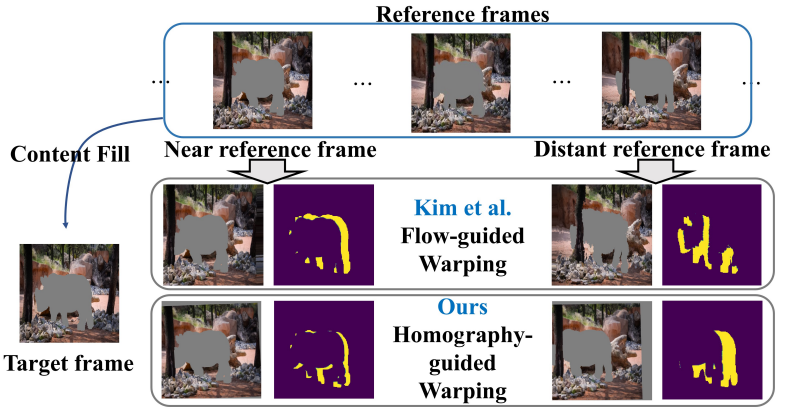

Figure 1: **Comparison of two different warping methods.** The yellow region indicates the part in the target frame that can be filled up by aligning the reference frame. Note that our approach successfully aligns a distant reference frame while the flow-based approach [12] fails.

Early works for video inpainting use a patch-based optimization technique [7, 8, 18, 26]. Among them, Huang *et al*. [8] proposed a global flow field-based optimization to preserve the temporal consistency, and show the state-of-the-art quality video results. However, the trade-off against the effectiveness is its limited practicality due to its intensive computational cost and vulnerability to noisy optical flows. Recently, two seminal works have proposed deep feed-forward methods for video inpainting [11, 12, 23]. Wang *et al*. [23] proposed CombCN by combining 3D and 2D CNNs, but their model works on low-resolution videos with fixed square holes, limiting its application to general video object removal. Kim *et al*. [12] proposed VINet, which aggregates information by flow warping from nearby frames to the target frame. However, its internal dependency on the optical flow restricts the size of the temporal search window, which sometimes leads to boundary artifacts and blurry textures inconsistent with global video contents.

To overcome the issues mentioned above, we propose a novel coarse-to-fine network for video inpainting (see Fig. 2). We use a set of sampled video frames as the reference to take visible contents to fill the hole of a target frame. Our proposed network consists of two stages. The first stage is an alignment module that uses computed homographies between the reference frames and the target frame. The visible patches are then aggregated to fill in the target holes roughly. Despite being able to model only global transformations (*e.g*., affine, perspective), we observe that homography based alignment provides a much larger temporal search window compared to the optical flow based counterpart (see Fig. 1). We also empirically confirmed in Table 2(b) (Kim *et al*. vs Ours (+align)). The second stage consists of an attention module that matches the generated patches with known reference patches in space and time non-locally, and a softmax function that temporally pools the most relevant patches. This refinement stage compensates real motions that cannot be modeled by previous global transformations. Both stages consist of large spatial-temporal window size for the reference and thus enable modeling long-range correlations between distant information and the hole regions. Therefore, even challenging scenes with large or slowly moving holes can be handled. Our network is also designed with a decoder to synthesize the content that are never visible throughout the video and a recurrent propagation stream of encouraging temporal consistency in video results. Finally, the video results produced by our model are globally and locally consistent in a temporal aspect. The local consistency is achieved by the

recurrent flow estimator which enforces each consecutive frame to have little flicker artifacts. Meanwhile, the global consistency is more attributed to the long-range temporal aggregation (align-and-attend). **Our contributions can be summarized as follows:**

- We design the effective homography-based alignment block to pick up more visible contents from distant reference frames (**Align**). We provide empirical evidence that the flow-based alignment strategy has a clear drawback in aligning distant information in reference frames compared to the proposed method.

- To compensate for the real motions that cannot be modeled by previous global transformation, we design the non-parametric attention based refinement (**Attend**). We reformulate the original self-attention formulation to enable matching between the target hole and reference non-hole regions in a pixel level.

- Putting all together, we present a novel deep video inpainting framework, which leverages long-range video frames via a coarse-to-fine two-stage algorithm (**Align-and-Attend**).

- We conduct extensive ablation experiments to verify our proposals. Our final model shows *near real-time* inference speed while significantly outperforming the state-of-the-art learning-based approaches.

2 Proposed Algorithm

Let $X_1^T := \{X_1, X_2, ..., X_T\}$ be a set of video frames with spatial-temporal holes, and $Y_1^T := \{Y_1, Y_2, ..., Y_T\}$ be the reconstructed ground truth frames. We aim to learn mappings $G : X_1^T \rightarrow Y_1^T$, such that the prediction $\hat{Y}_1^T$ be as close as possible to the ground truth video Y_1^T, while being plausible and consistent in space and time. This can also be formulated as a conditional video generation task [10, 14, 19, 24] where we estimate the conditional $p(Y_1^T|X_1^T)$. To simplify the problem, we base on a Markov assumption [10, 12, 19, 24] to factorize the conditional into a product form, such that the generation of the t-th target frame $\hat{Y}_t$ is dependent on 1) current input frame X_t, 2) two previous output frames $\hat{Y}_{t-2}^{t-1}$, and 3) a set of sampled reference frames R as:

$$p(\hat{Y_1^T}|X_1^T) = \prod_{t=1}^{T} p(\hat{Y}_t|X_t, \hat{Y}_{t-2}^{t-1}, R). \tag{1}$$

The main idea of our approach is to use a set of sampled reference frames, R, that contains a sufficiently large temporal search window so that the visible information in the window can be fetched to inpaint the target frame with the globally coherent contents. According to our preliminary experiments, we sample every 10-th frames in a video to construct R, and this provides a much broader temporal window size than previous approaches [12, 23]. Another important design is to enforce each prediction to be temporally coherent with the past predictions, $\hat{Y}_{t-2}^{t-1}$. With this recurrent pathway, our model runs in an auto-regressive manner. Our proposed method outperforms existing learning-based methods [12, 23], and performs on par with the optimization-based method [8] while running at a much faster speed.

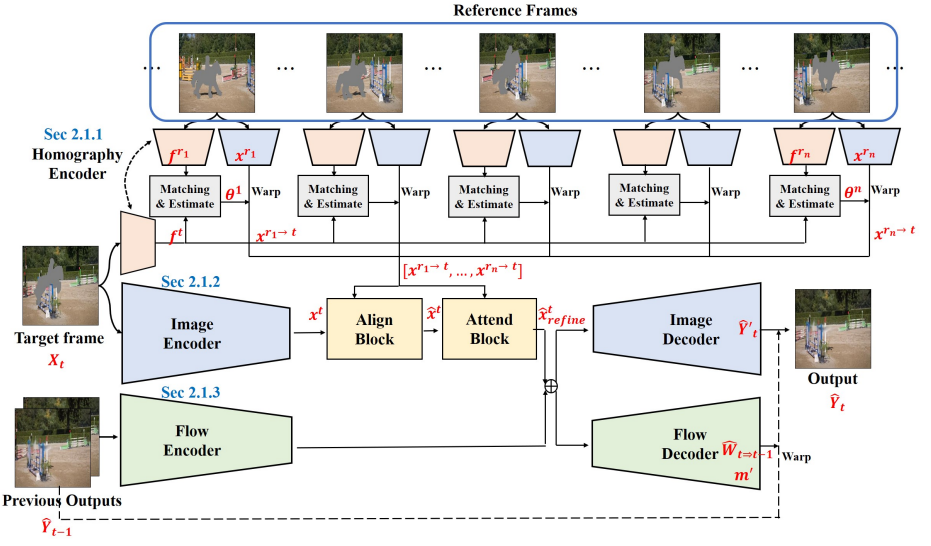

Figure 2: **The overview of the proposed Align-and-Attend network.** The whole model can be divided into three parts: 1) homography estimator, 2) align-and-attend video inpainter, and 3) flow estimator. First, we obtain affine transformation parameters using the homography estimator. Second, given the homography parameters, we accordingly align the reference frames. The following non-local attend block refines the initial coarse results. Finally, we enforce temporal consistency using the estimated flow.

2.1 Network Design

2.1.1 Homography estimator

Given a set of reference frames and a target frame, the goal of the homography estimator is to produce transformation parameters θ, which is to warp and align each reference frame onto the target frame.

Homography Encoder takes an image of size 256×256 pixels as input, and produces an embedded feature map $f \in \mathbb{R}^{c \times 32 \times 32}$, where c denotes the channel size . We use the same shared encoder for both the reference and target frames. We denote features of any reference frame by f^r, and those of the target frame by f^t.

Masked matching produces a measure of similarity between the reference and target feature maps. We denote the matching function as $M(f^r, f^t) := C$, such that $M : \mathbb{R}^{c \times 32 \times 32} \times \mathbb{R}^{c \times 32 \times 32} \rightarrow \mathbb{R}^{1024 \times 1024}$, where C is a cosine similarity map computed between channel-wise normalized f^r and f^t. We constrain the matching to happen only between the visible parts to deal with the holes regions. To this end, we use downsampled binary inpainting masks m^r and $m^t \in \mathbb{R}^{32 \times 32}$. With i and j denoting the spatial grid indices for f^r and f^t respectively, the correlation map is computed as:

$$C(i,j) = \begin{cases} f^r(i)^T f^t(j) & \text{if } m^r(i) m^t(j) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The similarity $C(i,j)$ is normalized by softmax over the spatial dimension of f^r, for each $f^t(i)$.

Transformation estimator takes the correlation map C as input and produces homography parameters θ between the reference and target. It is trained to output 6 parameters (affine transformation), such that $Te(C) := \theta$, and $Te : \mathbb{R}^{1024 \times 1024} \rightarrow \mathbb{R}^6$.

2.1.2 Align-and-Attend Video Inpainter

Our video inpainter is an encoder-decoder model consisting of the following components that are designed to reconstruct the target holes in a coarse-to-fine manner.

Image Encoder part follows the same architecture as in the homography estimator. Similarly, we denote encoded features of any reference frame by x^r, and those of the target frame by x^t, both of spatial size 32×32 pixels.

Alignment block is given the homography parameters computed as in Sec. 2.1.1, and accordingly aligns the reference feature maps onto the target feature map. We denote a reference feature map that is aligned to the target by $x^{r_i \to t}$, where $i \in [1...n]$, and n denote frame index and the number of reference frames, respectively.

After the alignment, the most relevant reference feature points are gathered that are in the spatial-temporal search window. To do so, we present an aggregation function that can evaluate the alignment for each reference feature maps and exclude irrelevant information such as newly introduced scene parts. Specifically, we measure the Euclidean distances (L2-norm) between each aligned reference frames and the target frame while ignoring the hole regions using the binary inpainting masks, $w^{r_i} = ||m^t m^{r_i \to t}(x^t - x^{r_i \to t})||_2$. The smaller the value of w^{r_i} represents better alignment of i-th reference frame. Therefore, the distance measure (i.e., $\frac{1}{w^{r_i}}$) is multiplied to the corresponding inpainting masks $m^{r_i \to t}$, which helps suppressing the low-quality aligned reference information. The following softmax across temporal dimension provides a volume, A^{r_i}. It weighs relevant pixels in the stack of $x^{r \to t}$ and flattens the stack into *one-frame* feature map as:

$$\hat{x}^r = \sum_i A^{r_i} x^{r_i \to t}, \quad A^{r_i} = \frac{exp(\frac{m^{r_i \to t}}{w^{r_i}})}{\sum_i exp(\frac{m^{r_i \to t}}{w^{r_i}})} \tag{3}$$

Finally, the initial coarse prediction of the target feature map $\hat{x}^t$ is then obtained as, $\hat{x}^t = (1 - \hat{m})x^t + \hat{m}\hat{x}^r$. Here, the $\hat{m}$ identifies the visible region that can be borrowed from the reference frames. In practice, it is obtained by max-pooling the visible regions (i.e., $(1 - m^t)m^{r_i \to t}$) along the temporal dimension.

Attend block is designed to model pixel-wise correspondences [25], *e.g.*, non-rigid motions, that cannot be covered by the previous global alignment stage. We propose to match the coarsely generated patches with the non-hole patches in the reference frame stack. We reformulate the self-attention formulation. Specifically, the target frame is embedded into the *query* while the reference frames are embedded into the *key* and *value*. In the module, the *query* features on the (roughly-filled) hole regions of target frame will be matched to every *key* feature on all non-hole regions of the reference frames. The query-key matching produces a spatio-temporal attention map which includes information about which pixel in which frame among the reference frames is important to refine the pixels in the target frame. Using this attention maps, we can finally pick up the most relevant *value* features on the non-hole region of the reference frames. Note we designed this module to be non-parametric, not requiring any embedding layers.

$$\begin{aligned} \hat{x}^t_{residual} &= softmax((\hat{x}^t_q)^T x^r_k)(x^r_v)^T, \\ \hat{x}^t_{refine} &= \begin{cases} \hat{x}^t_{residual} + \hat{x}^t & \text{if } m^r(i)\hat{m}(j) = 1 \\ \hat{x}^t & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

where $\hat{x}_q^t$ is a feature $\hat{x}^t$ is reshaped into a matrix with a shape of $\mathbb{R}^{c\times 1024}$. $\hat{x}_k^r$ and x_v^r are features that $x^{r_i\to t}, i\in[1...n]$ are reshaped into a matrix with a shape of $\mathbb{R}^{c\times 1024n}$. Only $\hat{x}_q^t$ and $\hat{x}_k^r$ are L2-normalized over channel axis for attention map computation. $x_{residual}^t$ is properly reshaped before summation with $\hat{x}^t$.

Image Decoder takes the final output of align-and-attend block, together with the warped features of the previous output frames $\hat{Y}_{t-2}^{t-1}$. Adding the intermediate representations from the previous time steps not only provides rich training signals to the whole network but also enhances the temporal coherency in video results. The decoder recovers the fine details for the hole regions to generate raw output, $\hat{Y'}_t$. It is designed with nearest-neighbor upsampling layers and successive convolutions to prevent checkerboard artifacts.

2.1.3 Optical flow estimator

Our optical flow estimator is a simple encoder-decoder model (same with the above image encoder-decoder). It computes flow fields between the previous output frame and the current target frame, which is used to enforce temporal consistency.

Flow Encoder takes the previous two output frames as input to propagate reusable information to the current time step. The encoded features are also fed into the decoder of the video inpainter.

Flow Decoder outputs optical flow from time step $t-1$ to t, and a composition mask. We use the predicted flow $\hat{W}_{t\Rightarrow t-1}$ to warp the previous output $\hat{Y}_{t-1}$ onto the current time step $\hat{Y'}_t$. We then blend the two frames into one by the estimated composition mask m' to obtain the final output of our whole network as, $\hat{Y}_t = (1-m')\odot\hat{Y'}_t + m'\odot\hat{W}_{t\Rightarrow t-1}(\hat{Y}_{t-1})$.

2.2 Objective functions

Homography estimation. We train the homography network using this objective function: $\mathcal{L}_{align} = (\frac{1}{s}\sum_{i=1}^{s}||G(\theta_{f^t}) - G(\theta_{f^{t*}})||_2) + (||\theta_{f^t} - \theta_{f^{t*}}||)$ The first part is introduced by [21], and the second part is the direct L1 loss of transformation parameters. $G(\theta_{f^t})$ and $G(\theta_{f^{t*}})$ correspond to the bilinear sampling grids that use predicted parameters and ground-truth respectively. Here, s denotes the total number of sampling coordinates. The homography estimation network is trained independently from the video inpainting network (inpainting network and the optical flow estimation network). After the training, we freeze the network's parameters and use it as a global affine transformer.

Video inpainting. The objective function is designed to capture pixel-wise reconstruction accuracy, perceptual similarity, and temporal consistency. The pixel-wise reconstruction loss is defined as follows: $\mathcal{L}_{hole} = \sum_t \left\|(1-m^t)\odot(\hat{Y}_t - Y_t)\right\|_1, \mathcal{L}_{valid} = \sum_t \left\|m^t\odot(\hat{Y}_t - Y_t)\right\|_1$ Here, t indexes over the number of recurrences, m_t is the binary mask, $\hat{Y}_t$ is the model output, Y_t is the ground truth, and $\odot$ indicates the element-wise multiplication. To ensure perceptual similarity between the predicted output and the ground truth, we adopt both image GAN loss, $\mathcal{L}_{im\_GAN}$, and video GAN loss, $\mathcal{L}_{vid\_GAN}$, that are introduced by [24]. For the temporal consistency, we use flow loss and warping loss which are defined as:

$$\mathcal{L}_{flow} = \sum_{t=2}^{T}(\left\|W_{t\Rightarrow t-1} - \hat{W}_{t\Rightarrow t-1}\right\|_1 + \left\|Y_t - \hat{W}_{t\Rightarrow t-1}(Y_{t-1})\right\|_1),$$
$$\mathcal{L}_{warp} = \sum_{t=2}^{T} M_{t\Rightarrow t-1}\left\|\hat{Y}_t - W_{t\Rightarrow t-1}(Y_{t-1})\right\|_1 \tag{5}$$

where $W_{t\Rightarrow t-1}$ is the pseudo-groundtruth backward flow between the target frames, Y_t and Y_{t-1}, extracted by FlowNet2 [9], $M_{t\Rightarrow t-1}$ is the binary occlusion mask [13]. Note that we use groundtruth target frames in the warping operation since the synthesizing ability is imperfect during training. We employ a curriculum learning scheme that increases the number of recursion, T, by 6 every 5 epochs. We increase the T up to 24.

The total loss is the weighted summation of all the loss functions, $\mathcal{L} = 100 \cdot \mathcal{L}_{hole} + 50 \cdot \mathcal{L}_{valid} + \mathcal{L}_{im\_GAN} + \mathcal{L}_{vid\_GAN} + 20 \cdot \mathcal{L}_{flow} + 20 \cdot \mathcal{L}_{warp}$.

3 Experiments

We evaluate our method both quantitatively and qualitatively. We compare our approach with state-of-the-art methods in three representative streams of study: deep image inpainting [29], deep video inpainting [12, 24], and optimization-based video inpainting [8]. Two metrics are mainly used for the evaluation. The first is the Inception score (FID) [24] extended to videos to measure the perceptual quality in spatio-temporal dimension. The second is flow warping errors between frames that measure temporal consistency of video results. We use public codes to obtain baseline scores.

3.1 Training

Homography estimation. We generate synthetic data using the Places2 image dataset [30]. Given a random image I_A, we generate the counterpart I_B by applying an arbitrary transformation to I_A. This provides us great flexibility to gather as many training data as needed, for any 2D geometric transformation. To simulate diverse hole shapes and sizes, we use the irregular mask dataset [17], which consists of random streaks and holes of arbitrary shapes. During training, we apply random affine transformations (*e.g.* translation, rotation, scaling, sheering) to the mask. All images are resized to 256×256 pixels for training.

Video inpainting. We employ a two-stage training scheme; 1) We first train the video inpainter without the alignment and the refinement stages to focus on learning a pure synthesis ability. To synthesize the training data, we follow the same protocol mentioned above. 2) We then add previously excluded stages along with the recurrence stream to the model. We fine-tune the whole model using videos in the Youtube-VOS dataset [27]. It is a large-scale video segmentation dataset containing 4000+ YouTube videos with 70+ various moving objects. Since the most realistic appearance and motion can be obtained from the foreground segmentation masks, we use them to synthesize the training video data. All video frames are resized to 256×256 pixels for training.

3.2 Testing

We use DAVIS dataset [20], which is widely used for video inpainting benchmarking. The videos are very challenging since they include dynamic scenes, complex camera movements, motion blur effects, and large occlusions. We obtain the inpainting mask by dilating the ground truth segmentation mask. Our method processes frame recursively in a sliding window manner.

3.3 User Study on Video Object Removal

We perform a user study to evaluate the visual quality of inpainted videos. We use 20 videos from the DAVIS dataset and compare our method with the strong baselines [8, 12]. A total of 25 users participated in this study. During each test, a user is shown video inpainting results by two different approaches, together with the input target video. We ask the user to check for both image quality and temporal coherency and to choose a better one. The users are allowed to play the videos multiple times to have enough time to distinguish the difference and make a careful judge. We report the ratio that each method outputs are preferred in Table 1(a). Our results are considered comparable to the [8], and much higher-quality than [12] by the human subjects. Note that our method runs faster than both approaches (see Table 1(b)). Some example results are shown in Fig. 3.

3.4 Quantitative comparison

We further compare our method with the baselines [8, 12, 24, 29] using both FID score and warping loss. Since we need the ground truth videos for this experiment, we composite target videos by overlaying foreground mask sequences extracted from other videos. To measure the FID score, we take 20 videos in the DAVIS dataset. For each video, we ensure to choose a different video out of the other 19 videos to make a mask sequence. We use the first 64 frames of both input and mask videos. To measure the flow warping errors, we use the Sintel dataset since it provides ground-truth optical flows. We take 32 frames each from 21 videos in the Sintel dataset and randomly select 21 videos of length 32+ from the DAVIS dataset to create corresponding mask sequences. For both metrics, we run five trials and average the scores over the videos and trials. We summarize the results in Table 1(c) and Table 1(d). We observe a similar tendency to the user study result.

To show the effectiveness of our model more concretely, we analyze the inpainting performance on two different settings: 1) fixed region inpainting ($H/4 \times W/4$ pixel area in the center) and 2) inpainting across different hole sizes. We used 30 DAVIS masks (avg area: 110^2) and the results are summarized in Table 2(a) and Table 2(b), respectively. The actual contents of hole regions are often discovered in reference frames, but previous flow-based methods are vulnerable at aligning frames with large displacements and occlusions [8, 12, 28]. Therefore they either produce inaccurate results for large holes or rely on heavy optimization, resulting in prohibitively slow speed. To overcome the limitation, we use the affine transform and the experiment (Table 2(b), Kim [12] v.s Ours (+align)) shows its effectiveness. Moreover, as demonstrated in Fig. 1, affine transformations can successfully align distant frames. Recent work also finds similar results [15]. Though, the homography-based alignment have difficulties in modeling complex and local motions. Our non-local attention based refinement stage is able to compensate for these errors by modeling pixel-wise correspondences, improving the first-stage coarse result significantly (Table 2(b), Ours (+align) vs. Ours (+align and refine)). Our final model outperforms [12] with a notable margin, especially when the hole becomes large (i.e., under more challenging conditions). The experimental results in Table 2 demonstrate the advantage of the proposed two-stage aggregation of long-range information in versatile application scenarios.

3.5 Ablation studies

We run an extensive ablation study to demonstrate the effectiveness of different components of our method. We measure the FID score and warping error following the same protocol as

Figure 3: Qualitative comparison with state-of-the-art video inpainting methods [8, 12] on DAVIS.

Human Preference Score	
Align-and-Attend Net (ours) / Huang *et al.* [8] / Tie	0.30 / 0.33 / 0.37
Align-and-Attend Net (ours) / Kim *et al.* [12] / Tie	0.58 / 0.19 / 0.23

(a) User study results.

Methods	time
Huang *et al.* [8]	several minutes per video
Kim *et al.* [12]	$\sim$ 12 fps
Ours	$\sim$ 15 fps

(b) Speed Comparison.

Methods	FID score
Yu *et al.* [25]	9.6948 ($\pm$ 0.708)
Huang *et al.* [8]	5.7032 ($\pm$ 1.158)
Wang *et al.* [24]	8.0130 ($\pm$ 0.701)
Kim *et al.* [12]	6.2852 ($\pm$ 1.769)
Ours	5.775 ($\pm$ 1.707)

(c) Spatial-Temporal Video Quality.

Methods	Warping error
Yu *et al.* [25]	0.0037 ($\pm$ 0.0003)
Huang *et al.* [8]	0.0017 ($\pm$ 0.0001)
Wang *et al.* [24]	0.0030 ($\pm$ 0.0001)
Kim *et al.* [12]	0.0019 ($\pm$ 0.0002)
Ours	0.0019 ($\pm$ 0.0001)

(d) Temporal consistency.

Table 1: Quantitative comparison with state-of-the-art video inpainting methods [8, 12, 25] on DAVIS.

in Sec. 3.4. The results are summarized in Table 3.

Network design choices. The main components of our network design are the two-stage feature aggregation part, together with the temporal propagation part. First, we investigate the importance of each stage in the align-and-attend network. If we drop the alignment stage out of the pathway, the refinement stage alone has to pick up valid reference patches to fill in the holes. However, it is difficult for the non-local module to match the *zero* patches to any reference patches without any priors. If we drop the refinement stage, the real video dynamics (*e.g.*, small, non-rigid motions) cannot be modeled, and the resulting videos would lack such fine details. To cancel out the effect of temporal propagation, we drop the flow estimator pathway. Without the recurrence, the temporal consistency is no longer well supported. If we remove multi-frame aggregation and the propagation parts, our network degenerates to a single image inpainting network. As shown in Table 3(a), all proposed components have complementary effects, and the best results are obtained when all components are fully used.

Masked matching in non-local attention module. In our non-local attention module, the coarsely completed region in the target hole is matched with the non-hole area in the reference frames. By doing so, the regions that remain as the holes are ignored during the refinement matching; Only those newly generated patches are touched during the refinement stage. To see the effectiveness of this matching method, we show the results when there was no such constraint (entire patches in the target frame is matched with the entire patches

Methods	FID score	warping error
Huang *et al.* [8]	6.8469	0.0020
Wang *et al.* [24]	9.4023	0.0031
Kim *et al.* [11]	7.5302	0.0023
Ours	6.9517	0.0022

(a) Experiments on fixed region inpainting. We used DAVIS and Sintel datasets.

Methods	FID score by hole size		
	small (area < 75^2)	medium (75^2 < area < 135^2)	large (area > 135^2)
Huang *et al.* [8]	2.1346	5.8417	6.7915
Wang *et al.* [24]	4.1827	8.6322	11.2268
Kim *et al.* [11]	2.2328	6.3428	7.6342
Ours (base)	3.1662	8.1035	9.4563
Ours (+align)	2.4551	6.0213	7.3230
Ours (+refine)	3.0151	6.9911	8.9018
Ours (+align and refine)	2.2175	5.9466	6.9132

(b) Different-sized holes inpainting experiments. **The lower the better.**

Table 2: Additional experiments

Multi-frame aggregation		Output Propagation	
Align	Refine	Flow estimator	FID score
			8.966 (±0.709)
✓			8.139 (±1.017)
	✓		8.577 (±0.838)
✓	✓		8.262 (±1.615)
		✓	7.515 (±0.608)
✓		✓	7.196 (±1.863)
	✓	✓	7.149 (±1.753)
✓	✓	✓	**5.775 (±1.707)**

(a) Ablations on network design.

Matching method	FID score
entire-entire	7.156 (±1.818)
hole-nonhole	**5.775 (±1.707)**

(b) Ablations on non-local matching methods.

Flow estimator	Warping error
	0.0027 (±0.0001)
✓	**0.0019 (±0.0001)**

(c) Ablations on recurrence stream.

Table 3: Results of ablation studies.

in the reference frames). As shown in Table 3(b), we observe that our proposed matching method is indeed useful, resulting in better video quality.

Recurrence stream. We report the flow warping errors to compare the temporal consistencies of video results before and after adding the recurrence stream (flow estimator pathway). As shown in Table 3(c), we observe the warping error is significantly reduced when there is a recurrence. This implies that propagating the previous output significantly improves the temporal consistency of videos. This is also consistent with the recent findings in [10, 11, 17].

The temporal window size of R. We sample every 10th frame in a video to ensure the reference frames contain a broad temporal context. In the meantime, we introduce a recurrence stream that allows the model to refer to the previous output frames, enabling the nearby information to be exploited simultaneously. We set the temporal stride of 10 empirically. We provide the FID scores of different temporal strides: 5 (5.903), 10 (5.775), 15 (5.869), 20 (6.028). The results show that temporal stride of 10 produces the best FID score.

The effect of Image and Video GAN loss. The image and video GAN loss ensure the spatial and temporal consistency of the outputs, respectively. If we drop any of them, we get inferior FID scores (from 5.775 to 5.925 and 6.012).

4 Conclusion

In this paper, we present a novel deep network for video inpainting. Our model fills in a target hole by referring multiple reference frames in a coarse-to-fine manner. First, we propose homography-based alignment between the reference and target frames to roughly inpaint the missing contents. Second, a non-local attention module refines the previously generated regions. Both stages provide large spatial-temporal window sizes that have not been achieved by existing flow-based methods. We validate the effectiveness of our approach in real object removal scenarios.

Acknowledgement

We thank Sunghoon Im and Jinsoo Leo Choi for helpful discussions.

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. In *IEEE Trans. Image Processing (TIP)*, volume 10, pages 1200–1211. IEEE, 2001.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[3] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. In *IEEE Trans. Image Processing (TIP)*, volume 12, pages 882–889. IEEE, 2003.

[4] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. In *ACM Trans. on Graph. (ToG)*, volume 31, pages 82–1. Citeseer, 2012.

[5] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001.

[6] 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 *Proc. of European Conf. on Computer Vision (ECCV)*, pages 682–695. Springer, 2012.

[7] Miguel Granados, James Tompkin, K Kim, Oliver Grau, Jan Kautz, and Christian Theobalt. How not to be seen—object removal from videos of crowded scenes. In *Computer Graphics Forum*, volume 31, pages 219–228. Wiley Online Library, 2012.

[8] 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.

[9] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017.

[10] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep blind video decaptioning by temporal aggregation and recurrence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4263–4272, 2019.

[11] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Recurrent temporal aggregation framework for deep video inpainting. In *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, volume 42, pages 1038–1052. IEEE, 2019.

[12] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

[14] Donghoon Lee, Tomas Pfister, and Ming-Hsuan Yang. Inserting videos into videos. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] 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.

[16] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, page 305. IEEE, 2003.

[17] 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 *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

[18] 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.

[19] Kwanyong Park, Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Preserving semantic and temporal consistency for unpaired video-to-video translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1248–1257, 2019.

[20] 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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016.

[21] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[23] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proc. of AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[24] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2018.

[25] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[26] Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time video completion. In *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, pages 120–127. IEEE, 2004.

[27] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

[28] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2019.

[29] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. In *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, volume 40, pages 1452–1464. IEEE, 2018.