Spatial-Temporal Residual Aggregation for High Resolution Video Inpainting

Vishnu Sanjay Ramiya Srinivasan[*1]
vishnusanjay.rs@gmail.com

Rui Ma[*2,1]
ruim@jlu.edu.cn

Qiang Tang[†1]
qiang.tang@huawei.com

Zili Yi[1]
yizili14@gmail.com

Zhan Xu[1]
zhan.xu@huawei.com

[1] Huawei Technologies
Canada

[2] Jilin University
China

Abstract

Recent learning-based inpainting algorithms have achieved compelling results for completing missing regions after removing undesired objects in videos. To maintain the temporal consistency among the frames, 3D spatial and temporal operations are often heavily used in the deep networks. However, these methods usually suffer from memory constraints and can only handle low resolution videos. We propose STRA-Net, a novel spatial-temporal residual aggregation framework for high resolution video inpainting. The key idea is to first learn and apply a spatial and temporal inpainting network on the downsampled low resolution videos. Then, we refine the low resolution results by aggregating the learned spatial and temporal image residuals (details) to the upsampled inpainted frames. Both the quantitative and qualitative evaluations show that we can produce more temporal-coherent and visually appealing results than the state-of-the-art methods on inpainting high resolution videos.

1 Introduction

Video inpainting is a long-studied task in computer vision and has wide applications such as video restoration [4], removing undesired objects [1, 18] etc. Although numerous efforts, including the traditional optimization based [2, 15, 23, 56] and the more recent deep learning based techniques [5, 6, 11, 20, 21, 24, 37, 43, 46], have been paid to tackle this challenge, most of the existing methods can only handle low resolution videos due to various reasons, e.g., memory or computation time constraints and the lack of high resolution training data.

In this paper, we aim for *high resolution* video inpainting without additional memory constraints and expect the inpainting network to be trained only on the low resolution videos. To this end, we propose STRA-Net, a novel spatial-temporal residual aggregation framework for high resolution video inpainting (Figure 1). The key idea is to first learn and apply a spatial-temporal inpainting network (STA-Net) on the downsampled low resolution videos. Then, we refine the low resolution results by aggregating the learned spatial and temporal *image residuals* (i.e., high frequency details) to the upsampled inpainted frames.

---

[*] Vishnu Sanjay Ramiya Srinivasan and Rui Ma are co-first authors.
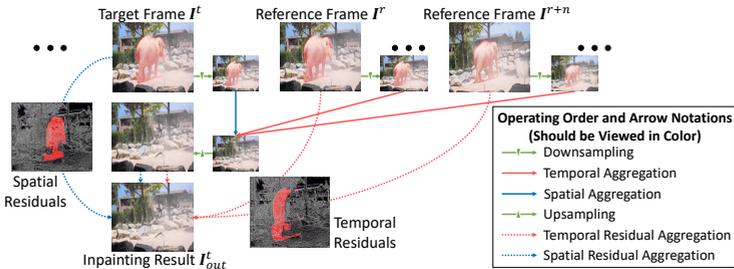[†] Qiang Tang is the corresponding author.

Figure 1: Given a video sequence (top row) with masks of undesired objects (red region), we first perform spatial and temporal aggregation on the downsampled low resolution frames. Then, we employ spatial and temporal residual aggregation for transferring residuals (details) to the upsampled impainted image and generate the high resolution inpainting result.

The STRA-Net is inspired by recent advance in high resolution image inpainting which employs the contextual (or spatial) residual aggregation [59] to recover the details of the inpainted region using the intra-frame information. Besides the spatial residual aggregation, we also introduce a novel *temporal residual aggregation* module to transfer image details from other neighboring frames. The STRA-Net is modularized, while the only trainable module STA-Net just needs to be trained and inferred on the low resolution videos. Hence, our method is memory efficient and not constrained by the resolution of the input video.

We conduct quantitative and qualitative comparisons with the state-of-the-art methods for high resolution video inpainting. From the experiments, our method can produce more temporal-coherent and visually appealing results. We also conduct ablation studies to verify the effectiveness of the proposed new modules. In summary, our contributions are as follows:

- We propose STRA-Net, a novel spatial-temporal residual aggregation framework that enables high resolution video inpainting without the memory constraints or requirements of high resolution training data.

- We combine a novel temporal residual aggregation module with the spatial residual aggregation to transfer high frequency image details from inter and intra video frames.

- We propose multiple new modules in STA-Net, such as a more robust joint image alignment module and the multi-scale spatial and temporal aggregation, to improve the performance of the learning-based video inpainting pipeline.

2 Related Works

Video inpainting is a long-studied yet still active research area [1, 18]. In this section, we mainly focus on the deep-learning based video inpainting techniques which have achieved compelling results recently. In addition, as a related topic, we summarize the latest methods for high resolution image inpaiting.

2.1 Deep Video Inpainting

A common solution for video inpainting is to complete the missing content using the spatial and temporal information gathered from the current or other video frames. In the deep learning regime, one typical practice for learning the spatial and temporal information in videos is to employ 3D convolutions to fuse the spatial-temporal features across multiple frames. Constrained by the GPU memory, inpainting methods employing 3D convolutions [5, 6, 14, 35] and recurrent propogation [9, 21] can only work on low resolution videos.

Another way to aggregate the spatial and temporal information is to compute the image alignment between frames and use the image correspondences to transfer information. Based on how the image alignment is estimated, most of recent techniques can be classified into either learning-based or flow-based methods. The learning-based methods [24, 25, 27, 29, 43] directly learn the alignment transformations from the pairwise image features. Afterwards, the missing regions can be predicted based on all aligned images, allowing information transferred even from temporal distant frames. However, such methods also often suffer from the memory constraints and require high resolution data for training. On the other hand, the flow-based methods [7, 11, 37, 44, 46] mainly perform pixel-wise data propagation between consecutive frames with the guidance of optical flow [16, 19]. The flow-based methods have shown promising results for inpainting high resolution videos, but they may produce artifacts when the flow estimation is not accurate, e.g., for videos containing fast or delicate motion. In addition, optical flow mostly detects image correspondences in the short temporal range and struggles with long temporal range pixel alignment [25]. Therefore, in this paper, we explore the possibility of combining strengths from both the learning-based and flow-based methods for a more robust joint image alignment. Besides performing the temporal aggregation using the proposed image alignment, we also employ the spatial aggregation which is proposed in the image inpainting [39, 41] to refine each inpainted frame.

2.2 High Resolution Image Inpainting

For video inpainting, there may occur some scenarios where no reference pixels can be found from other frames. An image inpainting step is usually incorporated into a video inpainting pipeline to fill the remaining missing regions with some visually plausible content. Recently, deep image inpainting methods [17, 30, 38, 39, 41, 42] have achieved impressive results by using techniques such as GAN [12] and attention based content transfer. For example, Yu et al. [41] proposed a contextual attention based inpainting method which can transfer features from distant spatial locations outside the hole region. In our spatial aggregation module, we use multi-scale contextual attention transfer to further refine the leftover region in which the pixels are synthesized by the network instead of copied from the reference frames.

While most of recent methods work effectively for low resolution images, they don't generalize well to high resolution images due to memory constraints or the requirement of high resolution training data. To overcome these limitations, Yi et al. [39] introduced the contextual residual aggregation (CRA) based algorithm HiFill to inpaint ultra high resolution images with less memory usage. Image residuals, which encode the high frequency details in the image, have been widely studied in low-level vision tasks, such as Laplacian pyramids construction [4, 34], edge detection [33] and image quality assessment [8]. In [39], the residual images computed by a simple yet effective downsampling and upsampling process are aggregated using the contextual attention transfer and then added to the low resolution inpainting result. In the end, a sharp high resolution inpainted image can be generated. In STRA-Net, we extend the idea of CRA to the temporal domain and transfer residuals from multiple temporal reference images to improve the low resolution video inpainting result.

3 Method

3.1 Overview of STRA-Net

The overview of our spatial-temporal residual aggregation framework (STRA-Net) is shown in Figure 2. The system takes a high resolution video with the masks of undesired regions
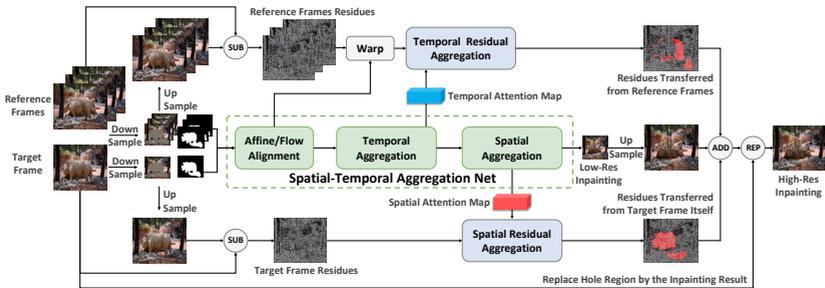
Figure 2: Overview of the STRA-Net. The STA-Net (highlighted by green) learns to inpaint the downsampled low resolution videos. The high frequency image details are added to the upsampled low resolution results by the spatial and temporal residual aggregation.

(holes) as input and outputs an inpainted high resolution video with the holes filled. The video is processed frame-by-frame in the temporal order. We refer the frame to be filled as the *target frame* and other frames as the *reference frames*. First, all the raw frames and masks are padded and downsampled to a lower resolution. Then, we learn a spatial-temporal aggregation network (STA-Net) to inpaint the low resolution target frame. To generate the residual images which contain the high frequency details, we upsample the previously downsampled frames and masks, and subtract them from the raw inputs. In the end, the spatial and temporal residual aggregation modules use the spatial and temporal attention maps computed from the STA-Net to generate a sharp inpainted frame at the original resolution. The inpainted frame will subsequently be used as a reference frame, providing more information for filling the following frames.

3.2 STA-Net

The STA-Net aims to inpaint the target video frame using the spatial and temporal information from the reference frames. Due to the robustness and extensibility of the learning-based method, we build our network structure based on a general learning-based pipeline, Copy-Paste Network (CPNet) [24]. In this section, we first provide a summary of CPNet and then introduce the STA-Net modules which improve the CPNet in all aspects.

CPNet summary. CPNet is a two-stage pipeline which contains three modules. In the first stage, an Alignment network is trained to compute affine transformation to align the reference frames to the target frame. Pixel-wise $L1$ distance is minimized between the target frame and the aligned reference frame on regions where the pixels are *visible* (non-hole) in both the frames. In the second stage, an encoder-decoder structure is used to extract and aggregate the features of the target and reference frames (Copy network), and reconstruct the target frame from the aggregated feature (Paste network). A key component in the Copy network is a *context matching module* which computes the features that can be copied from the aligned reference frames to the target frame. First, a cosine similarity $C_{r,t}$ is calculated between the target feature map $\mathbf{F}^t$ and each aligned reference feature map $\mathbf{F}^{r \to t}$. Then, a *masked softmax* operation is applied to the computed feature similarities to generate a *saliency map* (attention scores or weights) which is used to aggregate the features from each $\mathbf{F}^{r \to t}$ to $\mathbf{F}^t$. In the Paste network, the target frame feature is concatenated with the aggregated reference features and fed to the decoder, which consists of several dilated convolution blocks and upsampling operations. The target frame is reconstructed in the decoder by either filling the holes if any aggregated reference is available or generating the pixels using dilated convolutions if no reference information is found in the previous step. Here,
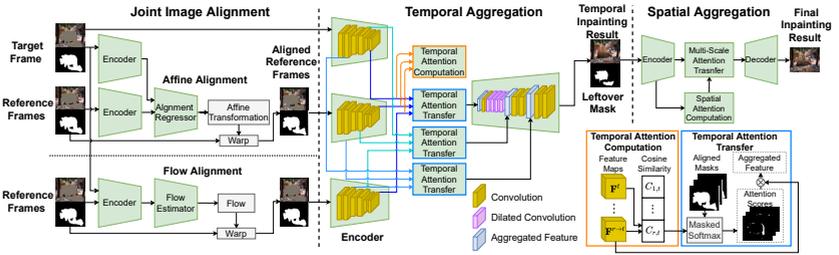
Figure 3: Overview of the STA-Net. All trainable modules are in green. The image encoders have the same architecture composed of Convolution and MaxPool layers, but their weights are only shared within the same module. Each Temporal Attention Transfer module (blue) uses the same attention scores computed from the Temporal Attention Computation module (orange) to transfer features at different scales, while we omit to draw the connections between them to avoid the clutteredness.

we only briefly summarize the pipeline of CPNet and we ignore the operations on the masks for the simplicity of presentation. More details about CPNet can be found in [24].

STA-Net pipeline. We extend the CPNet to three stages and propose novel improvements on each stage (Figure 3). In the first stage, we propose a joint alignment network which combines affine transformation and optical flow for better image alignment between the target and the reference frames. Secondly, we apply a multi-scale temporal aggregation network to copy information from references frames and generate an initial inpaiting result. Finally, we augment the CPNet with a new refinement stage which uses a spatial aggregation network to refine the result using contextual information from the target image itself.

Joint alignment network. Computing a robust image alignment within a proper temporal window is crucial to copy valuable and relevant contents from the other frames. The affine transformation based alignment [24], which we refer to as *affine alignment*, provides a large temporal window to copy information from distant frames and is best suited for static background based on our experiments. In contrast, the optical flow based alignment [1, 57] (*flow alignment* in short) has a small temporal range, but can handle local motion due to the dense image correspondences computed. In our approach, we jointly employ the affine and flow alignment to generate two types of aligned reference frames (Figure 3 left): one is computed by applying the learned affine transformation to each reference frame; the other is obtained by using the optical flow to propagate the pixels from the neighboring reference frames to the target frame. The affine transformation branch is implemented similarly as in CPNet, while for the optical flow branch, we directly use a pre-trained PWC-IRR network [16] to compute the optical flow. The target frame and all two types of aligned references frames are passed to the temporal aggregation module.

Multi-scale temporal aggregation network. After alignment, the target frame and aligned reference images are converted into multi-scale feature maps using a shared image encoder. Motivated by the multi-scale attention transfer and score sharing for efficient spatial feature aggregation introduced in [39], we propose *multi-scale temporal attention transfer* to aggregate the temporal features in a multi-scale manner. This module can also be treated as a multi-scale extension of the contextual matching module in CPNet. To increase the efficiency, we share the temporal attention scores calculated at the lowest resolution of the multi-scale feature maps and perform attention transfer on multiple higher resolution feature maps (Figure 3 middle). The attention scores are computed by applying a masked softmax function on the cosine similarity $C_{r,t}$ between the target feature $\mathbf{F}^t$ and the aligned reference

frame feature $\mathbf{F}^{r \to t}$ (Figure 3 bottom right). This multi-scale feature aggregation scheme can ensure more refined feature selection when transferring the features from reference frames. The temporally aggregated feature is then passed to a similar image decoder as in CPNet which outputs an inpainted image. Note that for certain regions in the result, there may be no feature transferred from other frames. Such regions correspond to a *leftover mask* where the temporal attention scores are zero for all reference frames. The pixels in the leftover mask are generated by the dilated convolutions in the decoder and thus may look blurry. Next, we will refine the regions covered by the leftover mask using the spatial aggregation.

Multi-scale spatial aggregation network. Given an initial inpainting result and a leftover mask, our goal is to refine the masked regions using the contextual information within the target image. This task is actually similar to the detail refinement step in the existing image inpainting methods. Here, we employ the refinement network from HiFill [39] as our multi-scale spatial aggregation module due to its efficiency and the ability to transfer the multi-scale features. The spatial aggregation network has a similar encoder-decoder structure to the temporal aggregation network, but operates in the spatial domain to copy multi-scale features from the contextual image patches in the target frame rather than the reference frames (Figure 3 top right). The spatial attention scores are computed based on pairs of contextual patches which contain one patch from the masked region (holes) and one patch from the non-hole regions. Similar to the multi-scale temporal attention transfer, the spatial attention transfer also aggregates the target image features at multiple resolutions using the shared spatial attention scores. At last, the spatially aggregated feature is passed to the decoder which generates a refined inpainting result for the region covered by the leftover mask.

3.3 Spatial-Temporal Residual Aggregation (STRA)

The STA-Net can work well for inpainting low resolution videos, but may not handle high resolution videos due to the memory constraints. One naive way to use STA-Net for high resolution video inpainting is to first downsample the input video and then upsample the inpainted result of STA-Net. However, this process will generate blurry results due to the upsampling. Inspired by the contextual residual aggregation based algorithm HiFill for efficient high resolution image inpainting [39], we learn and apply STA-Net at the low resolution and then perform spatial and temporal residual aggregation to improve the inpainting result (Figure 2). We extract the temporal and spatial residuals from the reference frames and the target frame by a regular subtraction process described earlier. Then, the affine and flow alignment learned from the STA-Net are used to align the temporal residual images to the target frame. Next, we perform temporal residual aggregation by using the temporal attention scores computed in the temporal aggregation network to transfer high frequency details from aligned temporal residual images to the upsampled STA-Net output. Similarly, we perform spatial residual aggregation to transfer details from surrounding patches to further refine the inpainted region as in HiFill. In the end, we succeed to generate a sharp inpainted output at its original resolution.

4 Training

The only trainable module in our framework is the STA-Net which is trained end-to-end by randomly sampling five frames from a synthesized training dataset (at 512x512 resolution) and optimized with the losses proposed below. In the inference stage, the network can take more than five frames (e.g., 20 in our implementation) to compute the image alginment and spatial-temporal aggregation across the frames.

Method	432 × 240			864 × 480			1296 × 720			1728 × 960		
	L1 ↓	SSIM ↑	PSNR ↑	L1 ↓	SSIM ↑	PSNR ↑	L1 ↓	SSIM ↑	PSNR ↑	L1 ↓	SSIM ↑	PSNR ↑
CPNet [24]	0.011	0.95	32.25	0.0125	0.93	32.40	0.014	0.92	31.69	0.018	0.89	30.20
STTN [13]	0.011	0.94	32.63	0.013	0.93	32.08	0.017	0.90	30.99	0.024	0.83	28.40
DFG [51]	0.008	0.94	33.12	0.008	0.95	33.27	0.008	0.96	**33.37**	0.009	0.95	32.37
FGVC [11]	0.006	0.95	33.20	0.006	0.95	33.20	0.006	**0.97**	33.20	0.006	**0.97**	**33.20**
Ours	**0.005**	**0.97**	**34.66**	**0.002**	**0.97**	**34.66**	**0.003**	**0.97**	33.14	**0.003**	**0.97**	33.14

Table 1: Quantitative comparisons on the Syn-DS$^+$ dataset at various resolutions.

Training dataset. To obtain videos with inpainting ground truth, we follow a similar procedure as in [24]. First, we composited various object/human masks to the 1.8M images from Places [45]. Then, we generated the image sequences (5 frames each) by random 2D image transformations. Besides the originally used masks from Pascal VOC [10], we also added masks from MS COCO [26] and Open Image [23] to increase the number and diversity of the missing region masks. In total, we have collected 335.7K masks. To further increase the dataset, we also downloaded natural scene videos from the Internet, processed them into 33.7K video clips (10 frames each) and composited the object masks similarly as before. All frames are resized and padded so that the final training data is at 512x512.

Losses. The total training loss for STA-Net is defined as follows:

$$L_{total} = 5 \cdot L_{align} + 10 \cdot L_{hole(visible)} + 20 \cdot L_{hole(leftover)} + 6 \cdot L_{non-hole}$$
$$+ 0.01 \cdot L_{perceptual} + 24 \cdot L_{style} + 1.2 \cdot L_{rec} + 0.001 \cdot L_{adv}. \quad (1)$$

Here, L_{total} is composed of two parts: one is similar to the losses defined in the CPNet for image alignment and temporal aggregation; the other employs the losses in HiFill to refine the leftover region by spatial aggregation. Specifically, L_{align} measures the $L1$ distance between the aligned reference frames and the target frame, and is used to supervise the alignment network. $L_{hole(visible)}$, $L_{hole(leftover)}$ and $L_{non-hole}$ are $L1$ distances measured on the reconstructed images for the pixels lying in the temporal aggregated, the leftover or the non-hole regions, respectively. Similar to CPNet, we apply perceptual and style losses, $L_{perceptual}$ and L_{style}, to further improve the visual quality of the results. At last, following HiFill, we use the $L1$ distance based reconstruction loss L_{rec} and the WGAN-GP loss [13] as the adversarial loss L_{adv} to train the spatial aggregation network. The mathematical definition and more training details can be found in the supplementary material.

5 Experiments

In this section, we conduct quantitative and qualitative evaluations as well as ablation studies of our method on various datasets. Our method produced superior results than the state-of-the-art methods for high resolution video inpainting in most cases.

Testing dataset. We test our method for inpainting various unseen videos after removing undesired objects or human specified by given masks. First, we chose 50 sequences from the commonly used DAVIS dataset [31] which provides high resolution videos with foreground object masks. As the original DAVIS object masks do not contain shadows, to make the inpainting result more realistic (without ghost shadows), we added the shadows to the object masks using our in-house annotation tool. We denote the shadow-annotated dataset as DS$^+$. All DS$^+$ videos are resized to 1080p (1920x1080). In addition, we collected 30 real world human-focused 1080p videos from the Internet and other datasets (denoted as HIN) and annotated some masks for undesired human in each video. To enable the quantitative evaluation, we also generated a synthetic dataset (Syn-DS$^+$) by imposing the DS$^+$ masks onto another 50 real world 1080p videos which contain various natural scenes and city views.

| | (a) | (b) | (c) | (d) |

Figure 4: Video inpainting results on the DS$^+$ dataset (1728x960): (a) Input frame with mask (b) STTN (c) FGVC (d) Ours.

	864 × 480				1728 × 960			
Method	CPNet [24]	STTN [43]	FGVC [11]	Ours	CPNet [24]	STTN [43]	FGVC [11]	Ours
Memory (GB)	3.3	5.7	2.6	**2.0**	10.1	20.6	18.7	**3.7**

Table 2: Comparison on the memory consumption of different methods.

5.1   Comparisons with Existing Methods

Baselines. We compare our STRA-Net[1] with the state-of-the art learning-based methods CPNet [24], STTN [43] and flow-based methods DFG [37], FGVC [11]. All experiments are done on a Tesla V100 GPU (32GB memory) using the pre-trained models downloaded from their websites. To verify the effectiveness of our different modules and the efficiency of our pipeline, we compare with the baselines on different video resolutions, ranging from low to high. Since the 1080p videos cannot be supported by certain baselines such as STTN on a single 32GB GPU due to the memory constraint, we resized the 1080p videos in our testing dataset to various lower resolutions. Proper padding is added to each resolution to fit the requirements of the baseline methods. Finally, we obtain four video resolutions as shown in Table 1. As our model is trained with 512x512 videos, after padding, only temporal and spatial aggregation (while no residual aggregation) are performed for the low resolution 432x240 videos. Hence, the results on the 432x240 videos can be used to evaluate the performance of STA-Net alone instead of the full STRA-Net. For all other resolutions, our STRA-Net will downsample the input video to 1/4 of the input resolution and run the STA-Net on the downsampled frames.

Quantitative comparison. As the Syn-DS$^+$ contains the ground truth frames after the synthetic DS$^+$ masks are removed, we conduct quantitative evaluation at each resolution of the resized Syn-DS$^+$ by computing the mean value of the standard L1, SSIM [52] and PSNR [22] metrics between the completed and ground truth frames. From Table 1, our method produces superior results than the learning-based and flow-based baselines in most cases.

To further verify the efficiency of our pipeline, we also compare the running time and memory consumption for each method. On average, to run on a 1728x960 frame, our STRA-Net takes 1.2 seconds which is similar to the CPNet (1s) and much faster than STTN (4s), DFG (48s) and FGVC (55s), showing the efficiency of our method. For memory consumption, we compare with the baselines at two resolutions and our full pipeline (with the residual aggregation enabled) reports the lowest memory in both cases (Table 2). It can also be observed that the memory of other methods grows much faster than ours since we can always downsample the input video to a lower resolution (e.g., 1/4 of input) for running the STA-Net. Note that the spatial and temporal residual aggregation can run on the CPU efficiently. Also, we didn't perform any memory optimization for all methods.

Qualitative comparison. We compare different methods on both real world videos in DS$^+$, HIN and the synthetic ones in Syn-DS$^+$. In Figure 4, our method produces more meaningful inpainting than other methods on DS$^+$. Due to the page limit, more results on HIN and

---

[1]Our code and the improved masks which include the shadow annotation on top of the original DAVIS masks are available at https://github.com/Ascend-Research/STRA_Net. The code has been tested on GPU (CUDA) and Huawei Ascend processor (CANN).
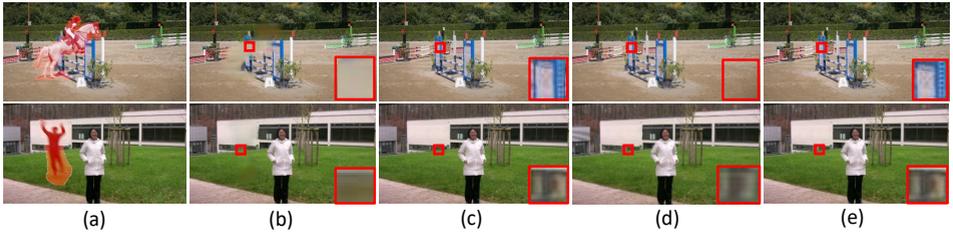
Figure 5: Ablation study for temporal modules on DS$^+$ (first row) and HIN (second row): (a) Input frame with mask (1080p) (b) Vanila temporal aggregation (270p) (c) Multi-scale temporal aggregation (270p) (d) After bilinear upsampling (1080p) (e) With temporal residuals added to (d).
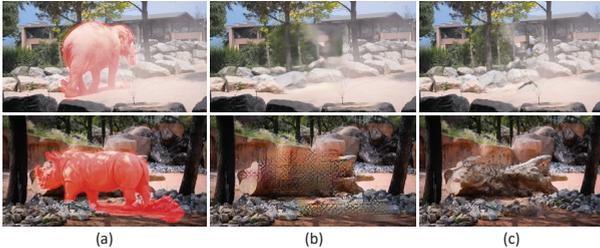

Figure 6: Ablation study for the spatial aggregation module on DS$^+$. (a) Input frame with mask (270p) (b) Without spatial aggregation (270p) (c) With spatial aggregation (270p).

Syn-DS$^+$ can be found in the supplementary material. Overall, our method can generate more temporally and spatially consistent results due to the proposed joint image alignment as well as the attention-guided multi-scale temporal and spatial aggregation.

User study. We conduct a user study by asking 15 volunteers from the computer science domain to rank the inpainting results from our method along with CPNet, STTN and FGVC. We randomly chose results (generated with 1728x960 input) of 20 sequences from DS$^+$ and 10 sequences from HIN to make a study of 30 tasks. From the study results, our method is ranked as the best in 39.8% cases, and the second best in 53.6% cases, comparing to the 54.9% and 30.0% of FGVC to be ranked as the best and the second. Note that for some results, the top two results are actually very similar. For these cases, we ask the user to randomly choose one as the best. Based on the user study, although our method got slightly inferior results than FGVC with some randomness, it performs far better than CPNet and STTN. Also, FGVC can take much longer time to compute the optical flow for high resolution videos, while our method runs much faster and requires much less memory since our STA-Net only runs on low resolution frames. More details of the user study setup and results are provided in the supplementary material.

5.2 Ablation Studies

We perform extensive ablation studies to identify the effectiveness of the proposed modules. As our method can naturally support high resolution videos, all ablation studies are conducted on 1080p videos unless the resolution is specified. The corresponding downsampled low resolution videos are 270p (480x270).

Analysis of different spatial and temporal modules. We conduct quantitative ablation studies for different spatial and temporal modules on the Syn-DS$^+$ dataset. The results in Table 3 show the contributions of individual components. We also visualize the effects of different temporal modules in Figure 5 on the real world DS$^+$ videos. The proposed multi-

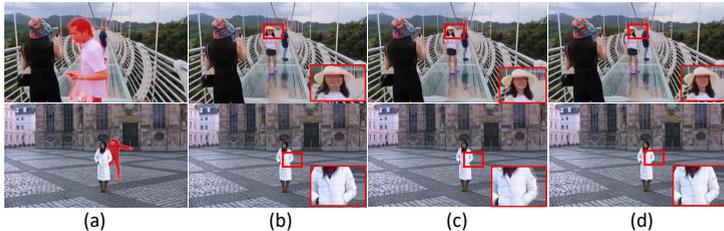|	(a)	(b)	(c)	(d)

Figure 7: Ablation study for different image alignment methods on HIN: (a) Input frame with mask (1080p); (b) Affine alignment only; (c) Flow alignment only; (d) Both. Flow alignment works better in the first row and affine alignment works better in the second. The joint alignment works well in both cases.

Module	432 × 240			Module	1728 × 960		
	L1 ↓	SSIM ↑	PSNR ↑		L1 ↓	SSIM ↑	PSNR ↑
Vanila temporal aggregation	0.024	0.85	28.36	Bilinear up-sampling	0.033	0.80	26.39
Multi-scale temporal aggregation	0.005	**0.97**	34.66	Temporal residual aggregation	0.005	0.96	33.11
+ Multi-scale spatial aggregation	**0.003**	**0.97**	**34.89**	+ Spatial residual aggregation	**0.003**	0.97	**33.14**

Table 3: Quantitative ablation study of different modules on Syn-DS$^+$. Left: multi-scale spatial and temporal aggregation modules on low resolution videos (no spatial or temporal residual aggregation is applied). Right: spatial and temporal residual aggregation modules on high resolution videos (both multi-scale temporal and spatial aggregation are enabled).

Alignment Method	L1 ↓	SSIM ↑	PSNR ↑
Affine transformation	0.005	**0.98**	32.05
Optical flow	0.017	0.927	29.21
Both	**0.003**	0.97	**33.14**

Table 4: Quantitative ablation study for different image alignment methods on Syn-DS$^+$. Results are generated at 1080p resolution.

scale temporal aggregation and the temporal residual aggregation can generally improve the details in the inpainting result comparing to the vanila baseline. In Figure 6, we can see the contribution of our multi-scale spatial aggregation for improving the inpainting quality on the central regions where no information has been transferred from other temporal reference frames, e.g., the region covered by the leftover mask. Overall, the results show our proposed modules are essential to improve the performance from the baselines.

Analysis of alignment method. We conduct an ablation study to evaluate the performance of different image alignment methods. As shown in Figure 7, affine alignment which can utilize global information and temporal features from far away frames produces better results if the mask region is small or if there is global motion like camera shake/pan involved. On the other hand, flow alignment works better when the mask region is large and the alignment can mainly be inferred based the local movements. Hence, both methods can excel in certain situations. Our final joint alignment module combines the strengths of both methods and achieves consistent results in different cases. In Table 4, quantitative results verify that our joint alignment method can generally improve the performance of each individual method.

6   Conclusion

We present STRA-Net, a novel and modularized learning-based framework for using spatial-temporal residual aggregation to achieve high resolution video inpainting. We believe the STRA-Net can be easily adapted to other video inpainting methods to improve their performance on high resolution videos. In the future, we would like to explore its application in other video processing tasks such as efficient HD video synthesis and animation [40].

7 Acknowledgement

References

[1] Anu Rachel Abraham, A Kethsy Prabhavathy, and J Devi Shree. A survey on video inpainting. *Int. J. Computer Application*, 56(9), 2012.

[2] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *IEEE CVPR*. IEEE, 2001.

[3] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.

[4] Rong-Chi Chang, Nick C Tang, and Chia Cheng Chao. Application of inpainting technology to video restoration. In *First IEEE Int. Conf. on Ubi-Media Computing*. IEEE, 2008.

[5] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proc. ICCV*, 2019.

[6] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Learnable gated temporal shift module for deep video inpainting". *BMVC*, 2019.

[7] Ya-Liang Chang, Zhe Yu Liu, and Winston Hsu. Vornet: Spatio-temporally consistent video inpainting for object removal. In *IEEE CVPR Workshops*, 2019.

[8] Manjusha Deshmukh and Udhav Bhosale. Image fusion and image quality assessment of fused images. *Int. J. Image Processing*, 4(5), 2010.

[9] Yifan Ding, Chuan Wang, Haibin Huang, Jiaming Liu, Jue Wang, and Liqiang Wang. Frame-recurrent video inpainting by robust optical flow inference. *arXiv preprint arXiv:1905.02882*, 2019.

[10] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Computer Vision*, 111(1), 2015.

[11] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *ECCV*, 2020.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[14] Yuan-Ting Hu, Heng Wang, Nicolas Ballas, Kristen Grauman, and Alexander G Schwing. Proposal-based video completion. 2020.

[15] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Trans. on Graph.*, 35(6), 2016.

[16] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *IEEE CVPR*, pages 5754–5763, 2019.

[17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. on Graph.*, 36(4), 2017.

[18] Shachar Ilan and Ariel Shamir. A survey on data-driven video completion. In *Computer Graphics Forum*, volume 34. Wiley Online Library, 2015.

[19] 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 *IEEE CVPR*, 2017.

[20] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *IEEE CVPR*, 2019.

[21] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Recurrent temporal aggregation framework for deep video inpainting. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 42(5), 2019.

[22] Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth International Workshop on Quality of Multimedia Experience*, pages 37–38. IEEE, 2012.

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[24] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proc. ICCV*, 2019.

[25] Ang Li, Shanshan Zhao, Xingjun Ma, Mingming Gong, Jianzhong Qi, Rui Zhang, Dacheng Tao, and Ramamohanarao Kotagiri. Short-term and long-term context aggregation network for video inpainting. In *ECCV*, pages 728–743, 2020.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[27] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Decoupled spatial-temporal transformer for video inpainting. *arXiv preprint arXiv:2104.06637*, 2021.

[28] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *Siam journal on imaging sciences*, 7 (4), 2014.

[29] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel net-works for deep video completion. In *Proc. ICCV*, 2019.

[30] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE CVPR*, 2016.

[31] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE CVPR*, 2016.

[32] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009.

[33] Mohsen Sharifi, Mahmood Fathy, and Maryam Tayefeh Mahmoudi. A classified and comparative study of edge detection algorithms. In *Proc. Int. Conf. on Information Technology: Coding and computing*. IEEE, 2002.

[34] Alexander Toet. Image fusion by a ratio of low-pass pyramid. *Pattern Recognition Letters*, 9(4):245–253, 1989.

[35] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *AAAI*, volume 33, 2019.

[36] Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time completion of video. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 29(3), 2007.

[37] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *IEEE CVPR*, 2019.

[38] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *IEEE CVPR*, 2017.

[39] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *IEEE CVPR*, 2020.

[40] Zili Yi, Qiang Tang, Vishnu Sanjay Ramiya Srinivasan, and Zhan Xu. Animating through warping: An efficient method for high-quality facial expression animation. In *Proc. ACM International Conference on Multimedia*, 2020.

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE CVPR*, 2018.

[42] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *IEEE CVPR*, pages 1486–1494, 2019.

[43] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020.

[44] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *Proc. ICCV*, 2019.

[45] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 40(6), 2017.

[46] Xueyan Zou, Linjie Yang, Ding Liu, and Yong Jae Lee. Progressive temporal feature alignment network for video inpainting. In *IEEE CVPR*, 2021.