

Spatial-Temporal Residual Aggregation for High Resolution Video Inpainting (Supplementary Material)

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 Yi¹
yizili14@gmail.com

Zhan Xu¹
zhan.xu@huawei.com

¹ Huawei Technologies
Canada

² Jilin University
China

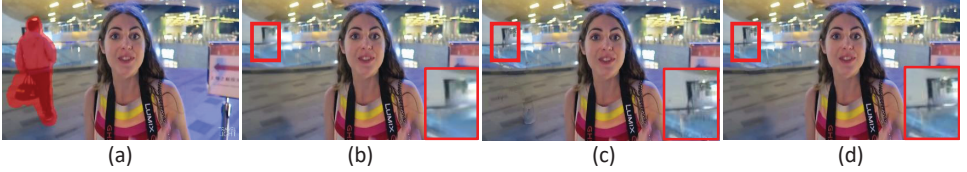


Figure 1: Video Inpainting result on human-focused real world HIN dataset: (a) Input frame with mask (1728x960) (b) STTN (c) FGVC (d) Ours.

1 Loss Functions

The definition of loss functions are similar as in CPNet [9] and HiFill [8]. The main difference is how the output of the main steps is calculated, e.g., we use the joint alignment while CPNet only uses the affine alignment. Also, our multi-scale spatial aggregation only applies on the leftover mask region instead of the full mask as in HiFill. Also, note that the loss functions are used to train the STA-Net which works on the downsampled low resolution frames. The specific definitions are as follows:

$$L_{align} = \frac{1}{N} \cdot \sum_t \sum_r \| (1 - M^t) \odot (1 - M^{r \rightarrow t}) \odot (X^t - X^{r \rightarrow t}) \|_1 \quad (1)$$

$$L_{hole(visible)} = \frac{1}{N} \cdot \sum_t \| M^t \odot C_{visible} \odot (Y^t - Y_{GT}^t) \|_1 \quad (2)$$

$$L_{hole(leftover)} = \frac{1}{N} \cdot \sum_t \| M_{leftover}^t \odot (Y^t - Y_{GT}^t) \|_1 \quad (3)$$

$$L_{non-hole} = \frac{1}{N} \cdot \sum_t \| (1 - M^t) \odot (Y^t - Y_{GT}^t) \|_1 \quad (4)$$

* Vishnu Sanjay Ramiya Srinivasan and Rui Ma are co-first authors.

[†] Qiang Tang is the corresponding author.

$$L_{perceptual} = \frac{1}{N} \cdot \sum_t \frac{1}{P} \cdot \sum_p \|\phi_p(Y_{comb}^t) - \phi_p(Y_{GT}^t)\|_1 \quad (5)$$

$$L_{style} = \frac{1}{N} \cdot \sum_t \frac{1}{P} \cdot \sum_p \|G_p^\phi(Y_{comb}^t) - G_p^\phi(Y_{GT}^t)\|_1 \quad (6)$$

$$L_{rec} = \frac{1}{N} \cdot \sum_t \|M_{leftover}^t \odot (\tilde{X}^t - X^t)\|_1 + \|(1 - M_{leftover}^t) \odot (\tilde{X}^t - X^t)\|_1 \quad (7)$$

$$L_{adv} = \frac{1}{N} \cdot \sum_t -\mathbb{E}_{\tilde{X}^t \in \mathbb{P}_g} [D(\tilde{X}^t)] \quad (8)$$

In above equations, X^t is the downsampled input frame at time t and M^t is its corresponding input mask. Y^t is the final inpainting result of X^t and Y_{GT}^t is the ground truth for Y^t . N is the number of frames in each sample of the training videos. For our current training data, $N = 5$. \odot is the element-wise multiplication. In Equation 1, $X^{r \rightarrow t}$ and $M^{r \rightarrow t}$ is the aligned reference frame and its aligned mask computed by aligning the reference frame X^r to X^t using the joint alignment module. In Equation 2 and 3, $C_{visible}$ is the aggregated temporal attention scores (see the main paper and [9] for the definition); $M_{leftover}^t$ is the leftover mask for frame t which is defined as $M_{leftover}^t = M^t \odot (1 - C_{visible})$. In Equation 5 and 6, Y_{comb}^t is the combination of the inpainting output Y^t in the hole region and the input X^t outside the hole; p is the index of the pooling layer in VGG-16 [9] and $\phi_p(\cdot)$ is the output of the corresponding layer; $G_p^\phi(\cdot)$ is the gram matrix multiplication [9]. In Equation 7 and 8, \tilde{X}^t is the generator output which is defined as $\tilde{X}^t = G(X^t, M_{leftover}^t)$, where $G(\cdot, \cdot)$ is the generator; $D(\cdot)$ is the discriminator and \mathbb{P}_g is the distribution of the input frames. Note that the adversarial training loss is designed in the same way as in [9]. More details about defining the WGAN-GP loss for the multi-scale spatial aggregation can also be found in [9].

2 More Training Details

In this section, we provide more details about training the STA-Net. The network is trained on the proposed synthetic training dataset which contains numerous short video clips and the ground truth object masks. We train the model on 2 Tesla V100 GPUs for 10 epochs with the batch size set to 40. In our experiments, we increase the weight of L_{align} so that the alignment module gets trained first and the later training converges faster. The joint alignment and temporal aggregation parts of the network are warmed-up for 2 epochs before adding the spatial aggregation module to the network for training stability reasons. The model is implemented in PyTorch.

3 More Experiment Results

Qualitative comparison. In Figure 1 and 2, we show more results for qualitative comparisons of our methods with the state-of-the-art learning and flow-based methods on the synthetic Syn-DS⁺ and real world HIN dataset. We also attach a demo video that compares results from different methods for video inpainting on 1080p videos from the DS⁺ [9] and the HIN dataset. Note that some baselines cannot directly run on 1080p videos due to memory constraint. Therefore, we run these baselines on 1792x960 resolution and then resize the output to 1080p. The improvements of our method upon others can be observed more clearly

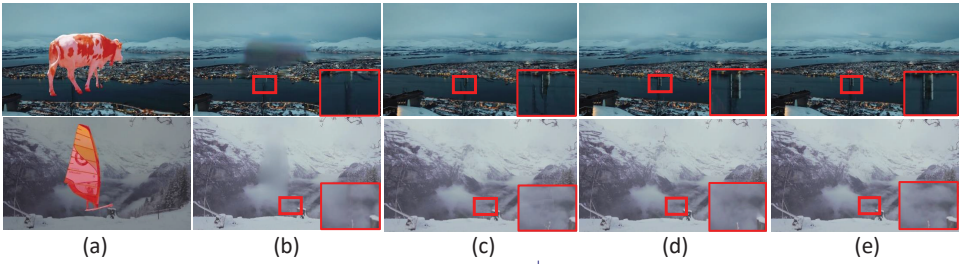


Figure 2: Video inpainting results on Syn-DS⁺ dataset: (a) Input frame with mask (1728x960) (b) STTN (c) FGVC (d) Ours (e) Ground Truth.

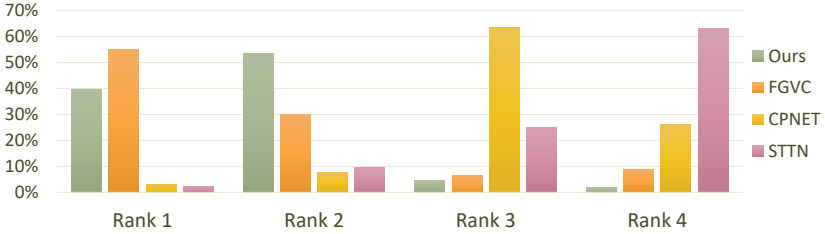


Figure 3: User study on 30 tasks composed of videos from DS⁺ and HIN. “Rank x” means the percentage of results from each model being chosen as the x-th best.

in the video. The results show that the learning based method STTN [14] tends to generate blurry inpaintings and the flow-based method FGVC [15] may produce artifacts near the mask boundaries. Our results are more temporally coherent and visually appealing.

User Study. In the user study, there are 30 tasks composed of video inpainting results generated at 1728x960 resolution of the DS⁺ and HIN dataset. In each task, the input video with the masks is first shown to the user, then results of the four methods are randomly arranged. The user is asked to rank the four results based on two criteria: 1) whether the inpainted videos look like real; 2) whether the inpainting results are spatially and temporally consistent. Figure 3 summarizes the user study results. It can be found that our results are comparable to the state-of-the-art flow-based method FGVC while outperform the learning-based methods CPNet [16] and STTN comfortably for the high resolution videos. Another important observation is our method has the lowest percentage of being ranked as the last which shows the robustness of our method.

References

- [1] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *ECCV*, 2020.
- [2] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
- [3] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proc. ICCV*, 2019.

- [4] 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.
- [5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. on Learning Representations (ICLR)*, 2015.
- [6] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *IEEE CVPR*, 2020.
- [7] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *ECCV*, 2020.