# An Attention-Based Video Inpainting Technique for Wire-Removal Scenarios

Jiacheng Hou
*Georgia Tech-Shenzhen*
Shenzhen, China
605680964@qq.com

Lixin Xu
*Georgia Tech-Shenzhen*
Shenzhen, China
lxu397@gatech.edu

Jianan Zhang
*Georgia Te ch-Shenzhen*
Shenzhen, China
1621544224@qq.com

Julong Li
*Georgia Tech-Shenzhen*
Shenzhen, China
jli3224@gatech.com

***Abstract*—In this paper, an attention-based video inpainting method is proposed for wire-removal scenarios, which is widely seen in the film making industry. In this model, the idea of onion-peel network and fuseformer algorithm are combined for better perceived and quantitative qualities. Besides, the model can also be used in other inpainting scenarios in addition to wire-removal. The results of PSNR and SSIM show that our modal achieves state-of-the-art quality.**

***Index Terms*—video inpainting, wire-removal, onion-peel, window-sliding**

## I. Introduction

Nowadays, many movies and TV series use the specially-made metal wire to help actors fly. In the post-processing stage, this kind of wire will be removed manually, which is extremely exhausting and trivial. Given this scenario, we aim to design an approach that can be used to finish the task naturally and efficiently.

Among varies of methods, video inpainting can be used for this purpose, since it is designed for image restoration from blocked masks. We can create blocked masks covering areas containing wires in every frames of video, and video inpainting method is able to infer the content without the wires.

## II. Related Works

In traditional methods, video inpainting generally draw experience from the image inpainting methods especially the patch-based image inpainting works which use a combination of searching and matching on a local or global scale to transfer known and reliable pixels to fill the unknown region in order to finish inpainting [1]–[3]. Later on, many researchers have explored and developed more optimized searching and matching techniques to yield better performance. However, for the video scene, applying image inpainting on videos frame by frame is not wise, cause it ignores the fact that a video is a sequence of images in time domain instead of a group of independent images. Thus, only considering the current time slot for inpainting may lead to artifacts and bad performance as a result. To overcome this issue, Weler et al. [4] construct a loss function to measure the time consistency of the inpainting result and adopt this kind of loss to the original reconstruction loss to optimize time consistency.

In recent years, deep learning has received a considerable amount of attention from the whole community due to its strong capacity in terms of high-level vision and semantic processing tasks. In video inpainting, deeper and deeper architecture arise to help model extracting more advanced features for the later processing. Pathak et al. [5] firstly utilize the deep auto-encoder architecture in inpainting videos and the adversarial loss to train the network. Followed by Lizuka et al. [6], a local and global discriminator scheme is applied in the original structure to generator more fine-grained textures. Apart from that, convolutional neural networks are also proved to be effective in video inpainting, Wang et al. [7] propose to combine 2D and 3D convolutional neural networks for inpainting videos, and specifically the 2D network is used to generate missing content and the 3D network is used to enhance the time consistency.

Since vision transformer has achieved SOTA performance in many vision tasks such as image classification, object detection and segmentation and so on. Some experts adopt the vision transformer in video inpainting task. Generally, the auto-encoder architecture will be remained, but between the encoder and decoder, various transformer models are added to help more precise content matching. In [8], a lightweight transformer and the onion-peeling algorithm are used to boost the performance. In [9], the feature map is divided into a number of patches with different scales and the matching is implemented on patches rather than the whole feature map to improve the texture restoration. In [10], a deep and stacked transformer model is proposed. Between each two adjacent transformer blocks, feature maps will be separated into many patches with overlapped regions and then fused, which is claimed to be effective for the deep model to learn sub-level token.

## III. Modelling

In our model, we propose to combine an improved version of onion-peeling algorithm [8] and the fuseformer [10] architecture to achieve better video inpainting. Besides, we propose a novel window-sliding technique to select information-rich frames for the reference of inpainting the target frame.

### A. Structure

As Fig. 1 denotes, we firstly utilize a symmetric encoder-decoder architecture to encode the input video to a feature representation and the frame size will be shrunk to one forth

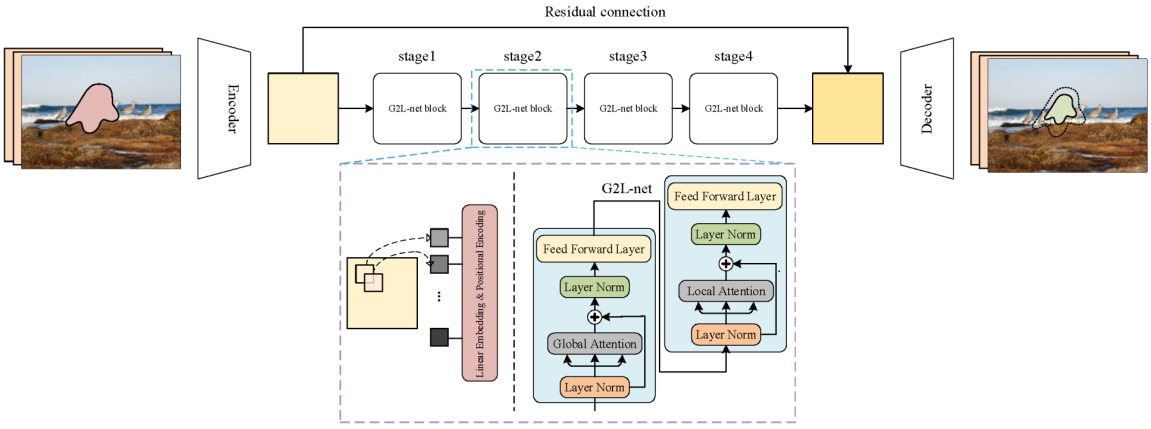



Fig. 1. Model Structure of Wire-removal Network

of the origin. The benefit of this operation is mainly that we can improve the processing efficiency of our model and reduce the memory consumption as well.

Then the inpainting step is implemented on the feature space of this video rather than on the video directly. The basic inpainting unit called G2L-net actually consists of two transformers in serial (the bottom blue part of Fig. 1). The structure of each transformer is similar and the only and core difference is that one transformer will focus on the global information and the other will focus on the local information. Four G2L-net blocks are connected in series, in which a decreasing size of receptive field is perceived.


Fig. 2. General to Local (G2L) technique

The idea of this design is that we can encapsulate global information through the transformer at first and then the focus is transferred to the locality of the feature space, which brings two benefits for the model. First, compared to the model that uses all global information, our model is absolutely more computation and memory saving, cause the local information is much easier to calculate than the global one. Second, this design follows a coarse-to-fine structure and is expected to make better use of the locality of the features. In image and video processing, the local information sometimes is very effective in inpainting task, cause local information usually is highly redundant, which means there is some repeated texture and details that can be used to enhance the inpainting quality.

### B. Improved Onion-peeling algorithm

Based on the pre-trained generative model, fuseformer, the missing content can be produced automatically, which is regarded as rough inpainting. Our work is to apply the onion-peeling algorithm on the synthesized content. The core idea of our algorithm is that we believe only the peel region of generated content is reliable and will be adopted to the corrupted video. For the remaining inpainting result, we will discard directly. The intuition for this design is simple, because the peel of missing region is spatially-closed to the known regions, thus this part has richer semantic information to reference and consequently is closer to the ground truth. Then, after several iterations, the missing region will be filled gradually until it diminishes.

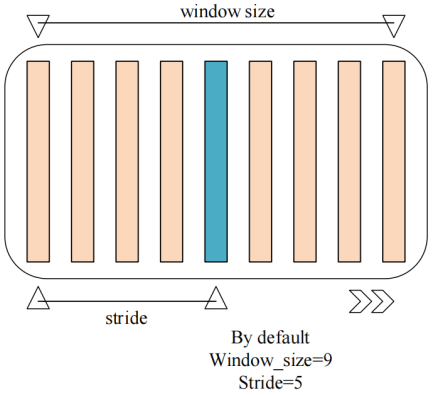



Fig. 3. Proposed window-sliding technique

### C. Window-sliding technique

One of the benefits of transformer is that it can process data with a parallel style, which means high-efficiency and better temporal consistency in video inpainting task. Therefore, in our inpainting generator, it receives a short clip from the whole corrupted video and restores this short clip to finish one inpainting iteration. However, the question is how to sample a short clip from the whole video. In previous works, random sampling is preferred, cause it can yield not bad result while very convenient to use. But, the core theory of video inpainting

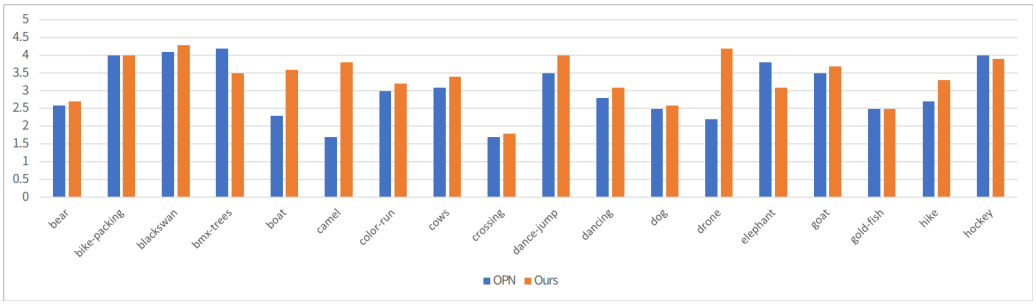



Fig. 4. Per-video Statistics on Subjective Perceived Quality

is to get missing content of current frame from frames in other time slots, so this scheme is seemed to be too simple and less designed for the task of video inpainting. For this reason, we propose an effective algorithm to sample reference through shifting a fixed size window in time domain. Within each window, a frame will be chosen for the reference frame, and the choosing strategy may come from computing the difference between this frame and the target frame. In this way, each shifting, the window will pick up a reference frame automatically and elaborately and a video clip can be formed. In this way, the synthesized video clip can be regard as a set of representative frames from the whole video, so it can help model to take patch matching and boost the performance.

## IV. Experiments

In this section, we test our method on video inpainting. We conducted our method using the test videos with a resolution of 424*240. To obtain the results of OPN method, we used the available official codes. We evaluated the method on the video completion task both qualitatively and quantitatively.

### A. Subjective Quality Assessment

We conducted a user study to subjectively compare our method against OPN video completion method. For the test video, we used video from DAVIS where every video frame has pixel-wise annotation for an object. We conducted the user study from our friends who are willing to participate. We collected the input and result of each video which can be shown to the participants, and they are asked to rank scores from 1 to 5.

Each test video was evaluated by 10 people, and the average score of OPN is 3.01 and Ours is 3.37. Per-video statistics of objects in DAVIS are show in Fig. 4, and the restoration of some samples are shown in Fig. 5.

Based on the average score, we can find that the scores are on the same level between our method and OPN. However, our method is approximately 2 seconds faster than OPN.

For the purpose of wire-removal, we collected our own dataset using a segment from a Chinese TV Drama. Because the wires are removed beforehand, we generate new wires and added to the inputs, as the first row of Fig. 6 shows. The masks of each frame, the performance of OPN, and the ground truth are shown as the second, third and fourth row respectively.

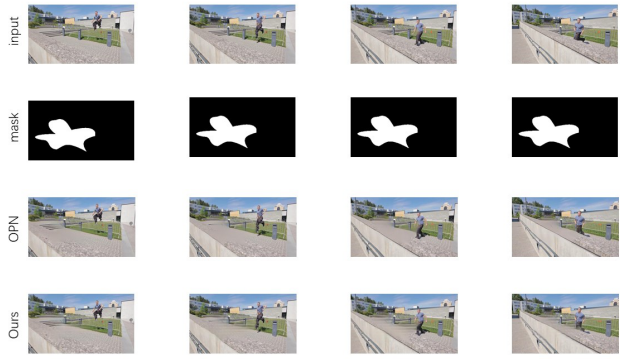



Fig. 5. Perceived Quality on a Sample from DAVIS

The performance of our method is demonstrated as the last row of Fig. 6. As can be seen, the perceived restoration quality is very similar, since the wires are very thin, and this is a very simple inpainting problem compared to those in the DAVIS dataset.


Fig. 6. Perceived Quality on Samples of a Wire-removal Scene

### B. Quantitative Quality Assessment

We have to acknowledge that it is hard to evaluate the mask removal since we can't find the ground truth, so that we synthesized imaginary objects on the top of existing videos in order to make test videos with known backgrounds. To achieve this goal, we shuffled the pairs of video and masks from DAVIS.

In Table I, we compare the PSNR and SSIM between our method and OPN method quantitatively, and our method shows the superiority over OPN method. We can easily find that PSNR is approximately 0.6dB over OPN method and SSIM is approximately 0.02dB over OPN method. Though the test videos are virtual, we still believe that this experiment can help us to understand the ability of two methods.

We believe that we achieve our goal by using new model which we add the optical flow in our method. Based on the progressive image processing, we trained a various of videos in different datasets and gain an ideal result which had a slide edge than OPN method.

TABLE I
PSNR AND SSIM SCORES IN WIRE-REMOVING SCENARIOS

| | PSNR | SSIM |
|---|---|---|
| OPN | 21.26 | 0.86 |
| Ours | 21.85 | 0.88 |

Besides, our model can also inpaint varies of objects in addition to wires. We tested both our method and OPN on different objects inpainting in metrics of PSNR and SSIM, and the results can be shown as Table II.

TABLE II
PERFORMANCE ON VARIES OF OBJECTS

| Objects | PSNR-Ours | SSIM-Ours | PSNR-OPN | SSIM-OPN |
|---|---|---|---|---|
| bear | 22.89 | 0.9046 | 22.08 | 0.8827 |
| bike-packing | 17.83 | 0.7744 | 17.51 | 0.7668 |
| boat | 25.61 | 0.9351 | 25.85 | 0.9264 |
| dog | 24.85 | 0.9218 | 23.17 | 0.9115 |
| drone | 18.3 | 0.7932 | 17.29 | 0.7647 |
| elephant | 22.9 | 0.8962 | 22.38 | 0.878 |
| goat | 26.03 | 0.934 | 25.25 | 0.9227 |
| gold-fish | 18.32 | 0.8214 | 17.94 | 0.8096 |
| hike | 19.63 | 0.9449 | 19.45 | 0.9271 |
| judo | 22.58 | 0.8909 | 21.98 | 0.8806 |
| koala | 21.8 | 0.8228 | 20.3 | 0.7905 |
| parkour | 24.29 | 0.9545 | 23.81 | 0.9395 |
| pigs | 15.72 | 0.7809 | 15.73 | 0.7579 |
| rhino | 20.03 | 0.8533 | 20.54 | 0.8419 |
| sheep | 21.36 | 0.9261 | 20.97 | 0.8912 |
| train | 19.56 | 0.8599 | 18.75 | 0.8316 |
| swing | 23.27 | 0.9261 | 23.23 | 0.9187 |

## V. CONCLUSION AND FUTURE WORK

In this paper, we come up with a model that handles wire-removal scenarios. Traditional attention-based methods tend to introduce blurriness because of the weighted average process, and therefore have impacts on the quality of restoration. In order to solve this issue, we adopt a more comprehensive and advanced mechanism to improve the restoration quality.

Based on the onion-peel network and fusion transformer, we develop a more advanced onion-peeling algorithm that can be applied on any generative model and has been proved to be effective. Besides, we propose a window-sliding algorithm to sample a information-rich video clip from the whole video to help inpainting.

Through testing on collected dataset with wires, as well as varies other datasets with masked areas, we assert that our model achieves better restoration quality in terms of PSNR and SSIM indices, which verifies our algorithms are effective.

For future work, we plan to construct a global to local architecture that can encapsulate global information and utilize the local information for inpainting to further exploit potential improvement in terms of visual quality.

## ACKNOWLEDGMENT

We sincerely thank the people of our team for their work in this project, from project proposal and data collection, to modeling and paper writing. During this whole process, we have had fruitful discussions and come up with many inspirational ideas.

We thank Jiacheng Hou for his leadership in this project. Without his ideas and work, we won't proceed in this direction. We thank Lixin Xu for his efforts in testing of algorithms, presentation, posters and paper writing. We thank Julong Li for his efforts in testing and experiment of algorithms. We thank Jianan Zhang for his efforts in mid-term presentation and the evaluation of algorithms.

We specially thank Professor AlRegib and TAs for their assistance and suggestions during this course.

## REFERENCES

[1] A. Efros and T. Leung. Texture synthesis by non-parametric sampling[C]// Proceedings of the Computer Vision and Pattern Recognition. Kerkyra, Greece, 1999: 1033 - 1038.

[2] M. Bertalmio, A. L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting[C]// Proceedings of the Computer Vision and Pattern Recognition. Kauai, USA, 2001: 355–362.

[3] Criminisi, A., Perez, P., Toyama, K. Region filling and object removal by exemplar-based image inpainting[J]. IEEE Transactions on Image Processing, 2004, 13(9).

[4] Wexler, Y., Shechtman, E., Irani, M. Space-time completion of video[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(3), 463–476.

[5] Pathak D, Krahenbuhl P, Donahue J, et al. Context encoders: feature learning by inpainting[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2536-2544.

[6] Lizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion[J]. ACM Transactions on Graphics, 2017, 36(4): Article No.107.

[7] Wang, C., Huang, H., Han, X., Wang, J. Video inpainting by jointly learning temporal structure and spatial details[C]// AAAI Conference on Artificial Intelligence.2019: 5232-5239.

[8] Oh S W, Lee S, Lee J Y, et al. Onion-peel networks for deep video completion[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4403-4412.

[9] Zeng Y, Fu J, Chao H. Learning joint spatial-temporal transformations for video inpainting[C]//European Conference on Computer Vision. Springer, Cham, 2020: 528-543.

[10] Liu R, Deng H, Huang Y, et al. Fuseformer: Fusing fine-grained information in transformers for video inpainting[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14040-14049.

# An Attention-Based Video Inpainting Technique for Wire-removal Scenarios

## ECE 6258 Group #6 Final Presentation

J. Hou, L. Xu, J. Zhang, and J. Li

Speaker: Lixin Xu

Date: Dec. 7 2021

Georgia Tech

# Problem Statement

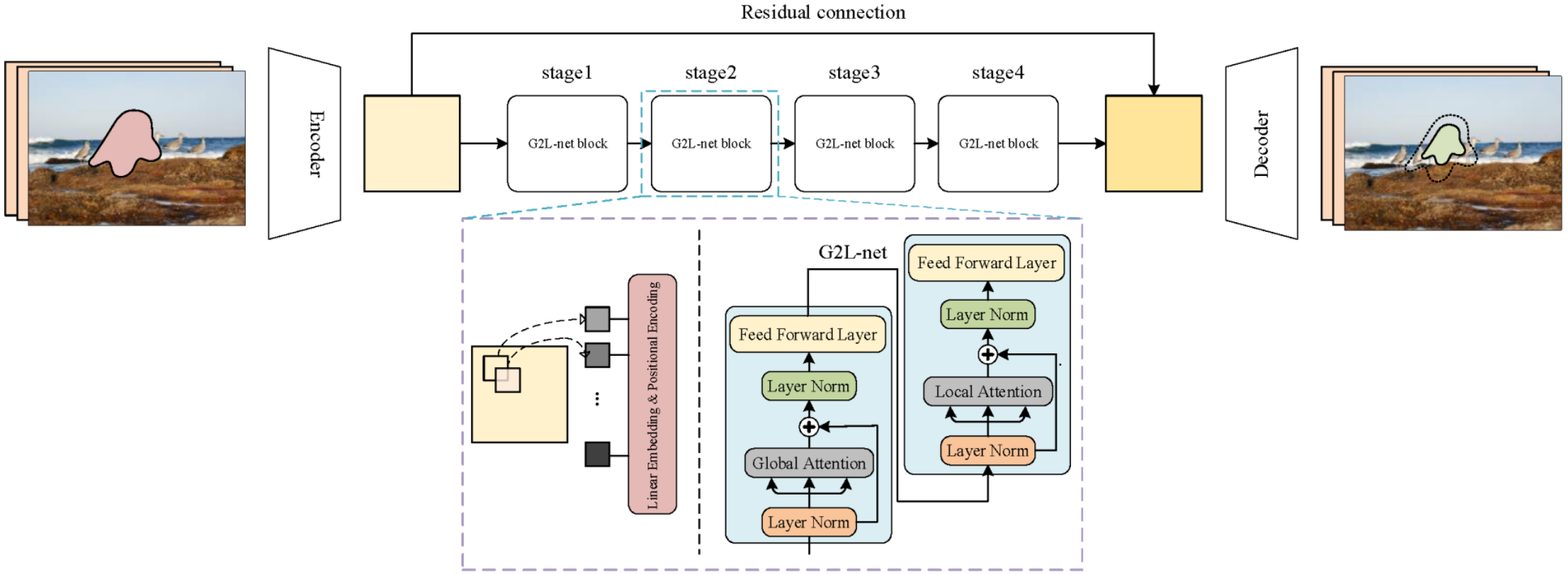
- In many movies and TV series, metal wires are used to help actors fly
- Develop an intelligent model to help people remove wires in the video
- Propose an **attention-based** video **inpainting** network for wire-removal

Georgia Tech

# Overall Methodology

- Datasets
- Encoder-decoder architecture
- G2L-net (Global to Local) with Transformer
- Window-sliding Technique
- Quality Assessment

# Proposed Model

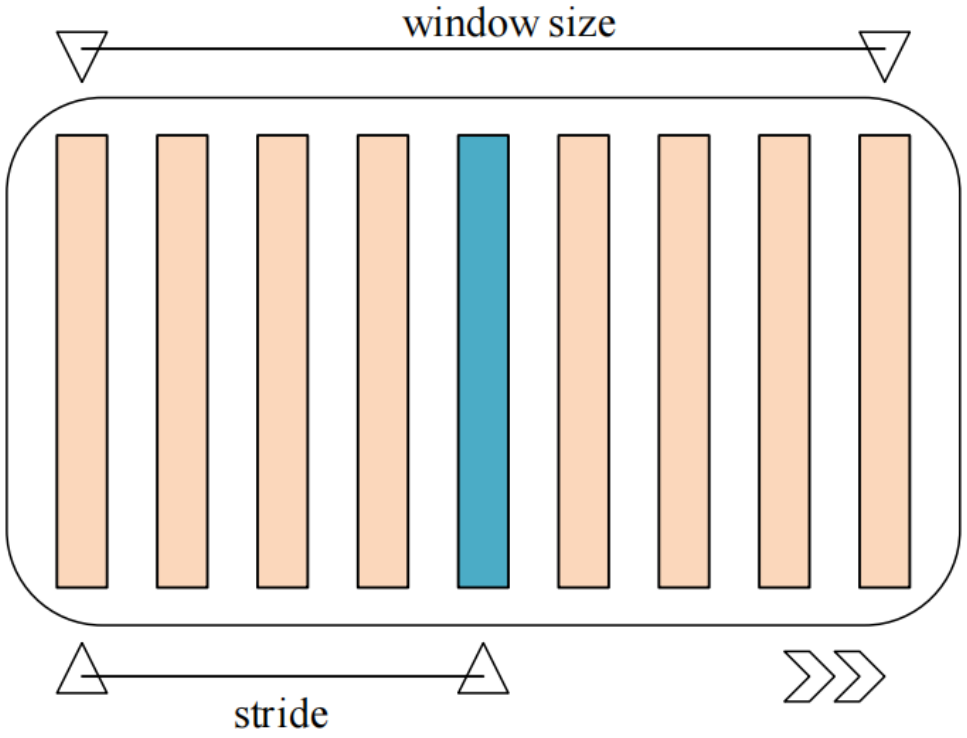


# G2L-net


- Composed of 2 Transformers
  - Global Attention
  - Local Attention
- Region of Attention, similar to human eyes

Georgia Tech

# Window-sliding technique




By default
Window_size=9
Stride=5

- Proposed technique to select frames with maximum difference
- Window size
- Stride
- Slide through time-series

Georgia Tech

# Improved Onion-peeling Algorithm

**Previous** [1]

- Inpaint a corrupted video frame by frame
- Each frame follows a progressive inpainting algorithm
- Completed frame will not be used for reference any more

**Ours**

- Inpaint a video clip by clip, which means a series of frames are inpainted in each iteration
- The completed video clip will be utilized in the next inpainting iteration.

[1] Oh S W, Lee S, Lee J Y, et al. Onion-peel networks for deep video completion[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4403-4412.

Georgia Tech

# Experiments

## Davis

## Collected

# Results Analysis

TABLE I
PSNR AND SSIM SCORES IN WIRE-REMOVING SCENARIOS

| | PSNR | SSIM |
|---|---|---|
| OPN | 21.26 | 0.86 |
| Ours | 21.85 | 0.88 |

TABLE II
PERFORMANCE ON VARIES OF OBJECTS

| Objects | PSNR-Ours | SSIM-Ours | PSNR-OPN | SSIM-OPN |
|---|---|---|---|---|
| bear | 22.89 | 0.9046 | 22.08 | 0.8827 |
| bike-packing | 17.83 | 0.7744 | 17.51 | 0.7668 |
| blackswan | 23.41 | 0.9035 | 22.66 | 0.8939 |
| bmx-trees | 27.99 | 0.9749 | 27.28 | 0.9652 |
| boat | 25.61 | 0.9351 | 25.85 | 0.9264 |
| goat | 26.03 | 0.934 | 25.25 | 0.9227 |
| gold-fish | 18.32 | 0.8214 | 17.94 | 0.8096 |
| hike | 19.63 | 0.9449 | 19.45 | 0.9271 |
| hockey | 22.43 | 0.9336 | 22.2 | 0.9254 |
| india | 23.65 | 0.888 | 22.48 | 0.865 |
| judo | 22.58 | 0.8909 | 21.98 | 0.8806 |
| koala | 21.8 | 0.8228 | 20.3 | 0.7905 |
| libby | 29.31 | 0.9679 | 28.63 | 0.9603 |
| train | 19.56 | 0.8599 | 18.75 | 0.8316 |
| swing | 23.27 | 0.9261 | 23.23 | 0.9187 |

Georgia Tech

# Conclusion

- Come up with model that handles wire-removal scenarios
- An improved version of onion-peeling algorithm
- Propose a window-sliding algorithm
- Testing of performance

Video link to our presentation:
https://drive.google.com/file/d/10WcyiTCV5Enb6q3EFI16kHMm8pIUQbjr/view?usp=sharing

Georgia Tech