Internal Video Inpainting by Implicit Long-range Propagation

Hao Ouyang* Tengfei Wang* Qifeng Chen
The Hong Kong University of Science and Technology

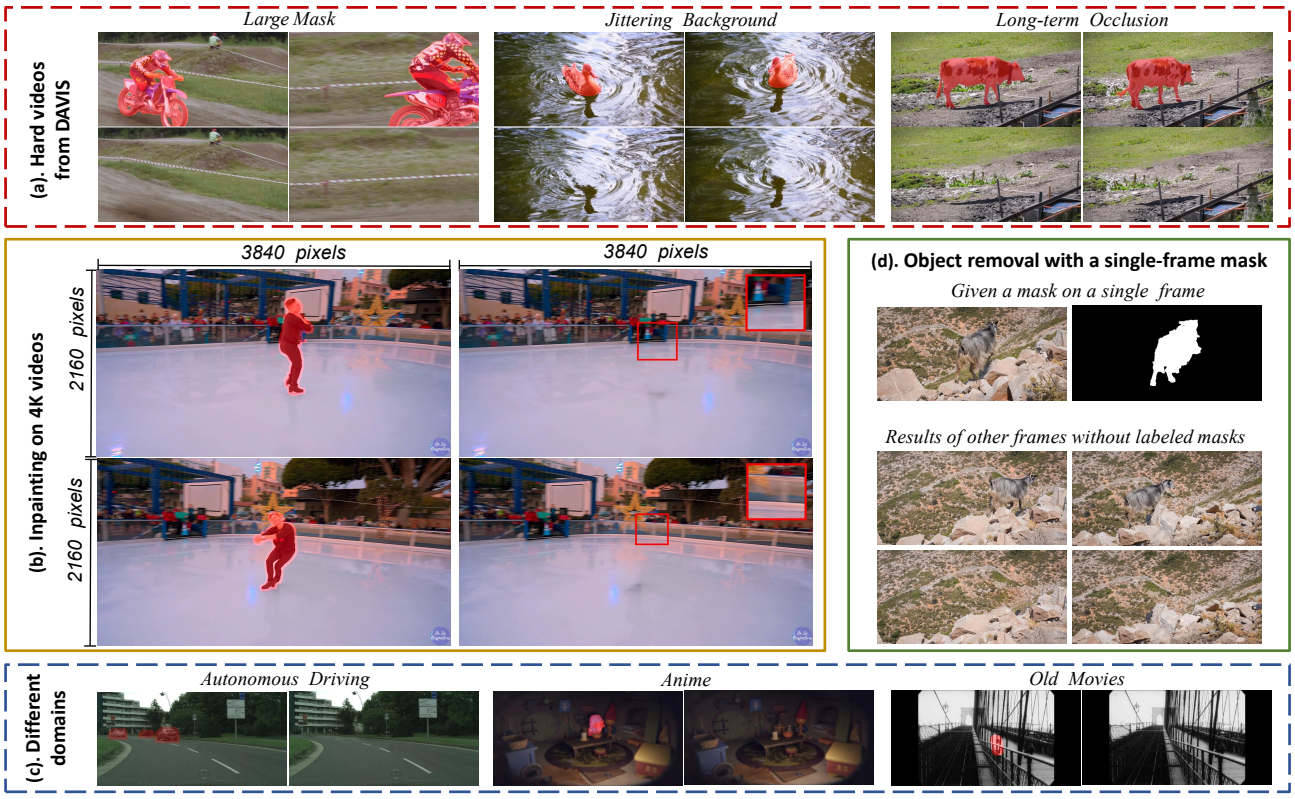

Figure 1: Ours results on (a) hard video sequences in DAVIS [29], (b) a 4K-resolution video, (c) videos from different domains [7], and (d) a video with a single-frame mask. Zoom in for details.

Abstract

We propose a novel framework for video inpainting by adopting an internal learning strategy. Unlike previous methods that use optical flow for cross-frame context propagation to inpaint unknown regions, we show that this can be achieved implicitly by fitting a convolutional neural network to known regions. Moreover, to handle challenging sequences with ambiguous backgrounds or long-term occlusion, we design two regularization terms to preserve high-frequency details and long-term temporal consistency. Extensive experiments on the DAVIS dataset demonstrate that the proposed method achieves state-of-the-art inpainting quality quantitatively and qualitatively. We further extend the proposed method to another challenging task: learning to remove an object from a video giving a single object mask in only one frame in a 4K video. Our source code is available at https://tengfei-wang.github.io/Implicit-Internal-Video-Inpainting/.

1. Introduction

Video inpainting is the problem of filling in missing regions in a video sequence with both spatial and temporal consistency. Video inpainting is beneficial for video editing, such as removing watermarks and unwanted objects. With the explosion of multimedia content in daily life, there

*Equal contribution

are growing needs for inpainting sequences from multiple domains and real-world high-resolution videos. It is also expected to alleviate the human workload of labor-intensive mask labeling for semi-automatic object removal.

The video inpainting task is still unsolved yet because the existing approaches cannot consistently produce visually pleasing inpainted videos with long-range consistency. Most traditional methods [25, 13, 12, 37] adopt patch-based optimization strategies. These methods have limited ability to capture complex motion or synthesize new content. Recent flow-guided methods [11, 39] propagate context information with the optical flow to achieve temporally consistent results. However, obtaining accurate optical flow in the missing region is non-trivial, especially when there is constantly blocked regions or the motion is complicated. Recent deep models [35, 15, 18, 27, 6, 20, 42] trained on large video datasets achieve more promising performance. However, the dataset collection process is time-consuming and labor-intensive, and these methods may suffer from performance drop when test videos are in different domains from training videos. Most recently, Zhang et al. [43] propose an internal learning approach to video inpainting, which avoids the domain gap problem as the model training is completed on the test video. While internal video inpainting is a promising direction, their approach sometimes generates incorrect or inconsistent results as this approach still depends on the externally-trained optical flow estimation to propagate context information. Therefore, we propose a new internal learning method for video inpainting that can overcome the aforementioned issues with implicit long-range propagation, as shown in Fig. 1.

Instead of propagating information across frames via explicit correspondences like optical flow, we show that the information propagation process can be implicitly addressed by the intrinsic properties of natural videos and convolutional neural networks. We will analyze these properties in detail and focus on handling two special challenging cases by imposing regularization. In the end, we manage to restore the missing region with cross-frame correlation and ensure temporal consistency by enforcing gradient constraints. We train a convolutional neural network on the test video so that the trained model can propagate the information on known pixels (not in masks) to the whole video.

We first evaluate the proposed method on the DAVIS dataset. The proposed approach achieves state-of-the-art performance both quantitatively and qualitatively, and the results of the user study indicate that our method is mostly preferred. We also apply our method to different video domains such as autonomous driving scenes, old films, and animations, and obtain promising results, as illustrated in Fig. 1 (c). In addition, our formulation possesses great flexibility to extend to more challenging settings: 1) A video with the mask on a single frame. We adopt a similar training strategy by switching the input and output in the above formulation to propagate masks. 2) Super high-resolution image sequences, such as a video with 4K resolution. We design a progressive learning scheme, and the finer scale demands an additional prior, which is the output from the coarse-scale. We show examples of the extended method in Fig. 1 (b) and Fig. 1 (d).

Our contribution can be summarized as follows:

- We propose an internal video inpainting method, which implicitly propagates information from the known regions to the unknown parts. Our method achieves state-of-the-art performance on DAVIS and can be applied to videos in various domains.
- We design two regularization terms to address the ambiguity and deficiency problems for challenging video sequences. The anti-ambiguity term benefits the details generation, and the gradient regularization term reduces temporal inconsistency.
- To the best of our knowledge, our approach is the first deep internal learning model that demonstrates the feasibility of removing the objects in a 4K video with only a single frame mask.

2. Related work

Image inpainting. Traditional image inpainting methods, including the diffusion-based [1, 2, 9] or patch-based [3] approaches, usually generate low-quality contents in complex scenes or with large inpainting masks. The recent development of deep learning has greatly improved inpainting performance. Pathak et al. [28] are the first to use the encoder-decoder network to extract features and reconstruct the missing region. The follow-up works [14, 40, 22, 41] improve the network designs to handle the free-form mask and adopted a coarse-to-fine structure [24, 21, 30, 36, 44] to use additional prior (e.g. edges and structures) as guidance. Our work utilizes a similar generation structure with [41] while adopts different training strategies and losses to improve the temporal consistency.

Video inpainting. Besides the spatial consistency, video inpainting involves another challenge, which is to ensure temporal consistency. Traditional methods usually complete regions by patch matching, including directly using 3D patches [37] or 2D patches with explicit homography-based or optical-flow-based constraints [25, 13, 12]. Most recent works utilize deep convolutional neural networks trained on a large video dataset to learn how to collect information from the reference frames to generate the missing contents. Several works use 3D CNN structures [6, 35] for feature extraction and content reconstruction but are extremely memory-consuming. Recent flow-based methods [11, 39] propagate temporal information to fill in the regions by optical flow or flow field and fill the remaining pixels with pre-trained image inpainting models. Other works

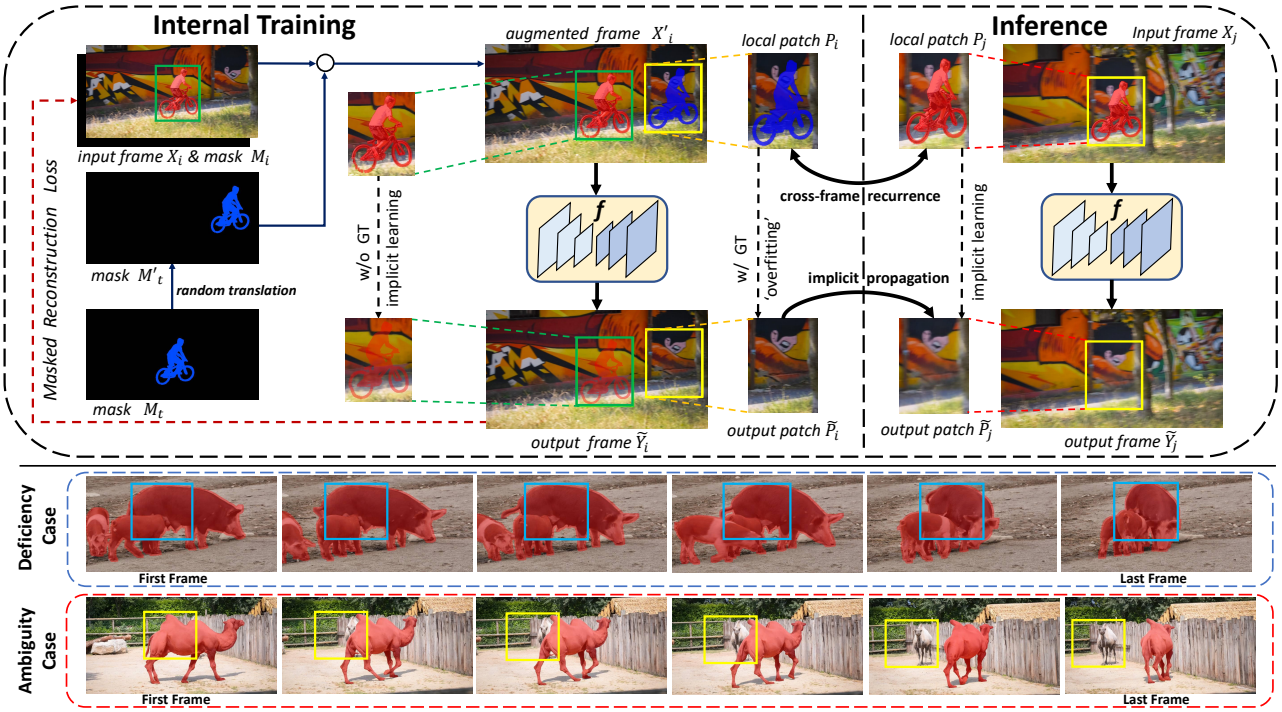

Figure 2: Overview of our internal video inpainting method. Without optical flow estimation and training on large datasets, we learn the implicit propagation via intrinsic properties of natural videos and neural network. By learning internally on augmented frames, the network f serves as a neural memory function for long-range information. When inference, cross-frame contextual information is implicitly propagated to complete masked regions. For non-ideal cases of deficiency and ambiguity where cross-frame information is unavailable or ambiguous, we design two regularization terms for perceptually-realistic and temporally-consistent reconstruction.

utilize 2D CNN to fuse information of target frames and reference frames with different context aggregation modules [15, 18, 27, 39, 20, 42]. Instead of learning on a large dataset, our approach explores another direction of internal learning in a video sequence, which is flexible and can be applied to different video categories.

Deep internal learning. Deep internal learning is a recent trending topic that has demonstrated great potential in image manipulation and generation. Ulyanov et al. [34] are the first to utilize deep model as prior on several image restoration tasks, including image inpainting. Internal learning with deep models has then been applied in various tasks such as image super-resolution [33], image generation [32, 31, 4], image inpainting [36], video motion transfer [5] and video segmentation [10]. The most related paper to our work is the deep internal learning in video inpainting [43] by jointing optimizing the generation of image and optical flow. However, their method relies heavily on the optical flow quality, which fails in large masks, slight motions, or incorrect frame selection. Our method adopts a totally different strategy by propagating the information from the known regions to the unknown regions implicitly by the neural network without optical flow and achieves more stable performance in the above-mentioned challenging cases.

3. Method

3.1. Problem formulation

Video inpainting takes a sequence of video frames $X : \{X_1, X_2, ..., X_N\}$ with the corresponding masks $M : \{M_1, M_2, ..., M_N\}$ of the corrupted region or the undesired objects, where N denotes the total number of the video frames. Our target is to generate the inpainted video $\tilde{Y} : \{\tilde{Y}_1, \tilde{Y}_2, ..., \tilde{Y}_N\}$ which is both spatially and temporally consistent. The known information can be represented as $\bar{X} : \{X_1 \odot (1 - M_1), X_2 \odot (1 - M_2), ..., X_N \odot (1 - M_N)\}$ and the quality of the inpainting depends profoundly on how we exploit the known information $\bar{X}$. We observe three interesting and inspiring properties:

1. **Cross-frame recurrence** Similar contents tend to appear multiple times in different frames in a video sequence. In Fig. 2, we use patch P_i in frame X_i and P_j in frame X_j to demonstrate this property. Note that P_i and P_j are not actually extracted in our pipeline but only to elaborate the properties. We denote the pixels of P_i excluding the masked region (blue region) as $\bar{P}_i$, the pixels of P_j excluding the masked region (red region) as $\bar{P}_j$. The upper part of Fig. 2 demonstrates an ideal case where the unmasked region $\bar{P}_j$ in inference

frame X_j also occurs in frame X_i where $\bar{P}_j \approx \bar{P}_i$.

2. **Neural network as a universal function** We can train a CNN to overfit to a certain amount of image data. With properly designed architecture, training an inpainting network f that generates high-quality inpainting results on $\bar{X}$ is achievable. As in Fig. 2, the network f generates nearly perfect content in the blue masked region in $\tilde{P}_i$ by overfitting the ground-truth.
3. **Translational equivalent convolution** Because of the weight sharing of convolution operation, if the input image of CNN is translated, the output of the network in each layer will translate in the same way [16]. This property indicates that if we regard $\bar{P}_j$ as a translated $\bar{P}_i$, the CNN f is supposed to generate high-quality contents for $\bar{P}_j$ as it for $\bar{P}_i$.

Based on these properties, we propose a novel view on internal video inpainting: by learning how to inpaint the known region, the model implicitly learns how to inpaint the unknown region. Instead of explicitly find correspondence using optical flow, the propagation process is implicitly learned by using an augmented mask. The detailed training strategy is as follows: we train a neural network f, which takes three inputs X_i, M_i, and M'_t in each training step, where i and t are randomly selected frame-index in the sequence. M'_t is augmented by M_t by random translation, and the input X_i is masked by the binary union of M_i and M'_t. Since the ground-truth pixels are not available in the unknown region masked by M_i, the reconstruction loss $\mathcal{L}_{rec}$ is only calculated in the known regions $\bar{X}_i$:

$$\tilde{Y}_i = f\left(X_i \odot (1 - M'_t \cup M_i)\right), \tag{1}$$

$$\mathcal{L}_{rec} = \|(\tilde{Y}_i - X_i) \odot (1 - M_i)\|_1. \tag{2}$$

As suggested by Yu et al. [40], we replace vanilla convolution layers with gated convolution layers. The detailed architecture of network f can be found in the **Appendix**.

3.2. Ambiguity and deficiency analysis

The solution with only the reconstruction loss works well for ideal cases where all three properties are satisfied (e.g. 'BMX-TREES' in DAVIS). However, in real applications, there are many exceptions. In Sec. 6, we further explore Property 2 on the relationship between the model capacity and the complexity of video sequence, and Property 3 on how the mask generation affects the inpainting quality. In this section, we focus on handling the violation cases on Property 1: the cross-frame recurrence.

We first investigate the ideal cross-frame recurrence in a video sequence. As shown in Fig. 2 upper parts, in the ideal case, there are consistent matches in the synthesized patches (P_i) and the inference patches (P_j). Two common non-ideal hard cases exist in the real video sequences: (a). There exist multiple unaligned matches and we call it ambiguity case. (b). There are no good matches and we call it deficiency case. We propose two regularization terms to improve the performance in these difficult situations.

3.3. Loss for ambiguity

Ambiguity exists when conflict matching patches occur in multiple frames. It usually appears when there are moving background objects or highly random textured regions in $\bar{X}$. Fig. 2 demonstrates a sample ambiguous sequence 'CAMEL' in DAVIS. The background camel is raising his head while the foreground camel passes by. For the local patch P_j in the first frame, there exist a set of patches $\{P_{i_1}, P_{i_2}, ...P_{i_n}\}$ with different contents (the position of camel face in this sequence). As expected, the predicted result of the moving region is blurry since it takes the average of all the ground truth, and we generate a blurry camel face shown in Fig. 8.

To address the ambiguity issue, We propose an ambiguity regularization term that brings the details back in the later training stage. This term is inspired by recent work [23], which matches blurry source areas with nearest neighbors in target regions. Let $\{s_p\}_{p\in P}$ and $\{t_q\}_{q\in Q}$ denote feature points collections of source image S and target image T extracted by a shared encoder. For each source point s_p, this loss tries to search for the most relevant target point under certain distance metric $\mathbb{D}_{(s_p,t_q)}$, and calculate $\delta_p(S,T) = \min_q \left(\mathbb{D}_{s_p,t_q}\right)$ [23]. In our case, we focus on how to utilize the cross-frame correlation within a video. For a output frame $\tilde{Y}_i$, we randomly select a frame X_t as the target, and improve the high-frequency details by imposing the anti-ambiguity regularization:

$$\mathcal{L}_{\text{ambiguity}} = \frac{1}{P}\sum_{p=1}^{P} \delta_p(\tilde{Y}_i \odot M_t, X_t \odot M_t). \tag{3}$$

3.4. Loss for deficiency

Deficiency situations exist when reliable cross-frame recurrence is not available. It usually appears when the foreground objects only have slight movement, and a large part of the video is always occluded in all frames. We also show a sample sequence 'PIGS' from DAVIS in Fig. 2, where a large region (e.g., the region in the blue mask) is constantly blocked. In this case, the network f cannot fill in these regions by directly propagating similar cross-frame information. However, as the model f is essentially a generative model learned on the internal video data, it can synthesize plausible new contents instead. The generated region follows the distribution of the known regions, which is acceptable in most cases. The quality depends on the complexity of the information (full results and the empirical analysis of extensive video sequences with deficient frames are presented in **Appendix**). As CNN f becomes a pure single

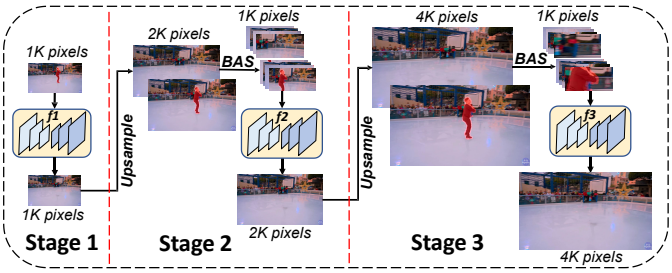

Figure 3: The pipeline of our progressive learning strategy for high-resolution video inpainting.

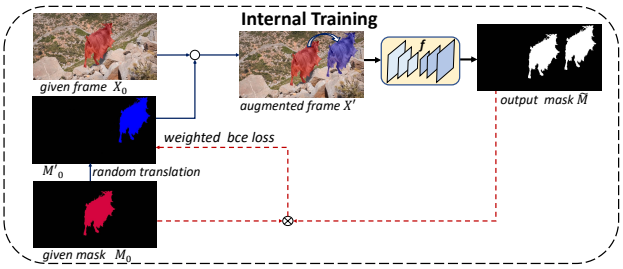

Figure 4: The training process of our mask propagation. Given only one mask, we can predict masks for the whole sequence.

frame generation model on the constantly blocked regions, severe flickering artifacts appear when there are continuous deficient frames.

The main issue is how to improve temporal consistency when encountering the deficiency cases. The recent works on improving the temporal consistency mostly depend on the pixel-to-pixel correspondence [17, 19, 8], while in the inpainting task, there is no such correspondence. We observe that even though the changes in the input surrounding region are trivial (such as small rotation), the network f can still generate a inconsistent result, which inspires us to apply constraints to the gradient of network f *w.r.t.* the input. Following settings in [8], suppose that the input X_i, M'_t, M_i, go through a slight modification g, where g is a random combination of homography transformation and image manipulation filters such as brightness change and blur kernel (when g is applied to mask, the manipulation filters are ignored.). The new inpainted result $\tilde{Y}'_i$ can be calculated as $f\{g(X_i) \odot (1 - g(M'_t) \cup g(M_i))\}$. We expect the change of outputs is consistent with the change of inputs [8], and calculate the gradient difference as:

$$\Delta_s = (\tilde{Y}'_i - \tilde{Y}_i) - (g(X_i) - X_i). \tag{4}$$

To minimize this term, we also need to exclude the pixels from unknown regions M_i and $g(M_i)$:

$$\mathcal{L}_{stabilize} = \|\Delta_s \odot (1 - M_i \cup g(M_i))\|_1. \tag{5}$$

The detailed ranges of each parameter used in g is attached in the **Appendix**. The overall training loss is the weighted sum of $\mathcal{L}_{rec}$, $\mathcal{L}_{ambiguity}$, and $\mathcal{L}_{stablize}$.

4. Extensions

4.1. Progressive high-resolution inpainting

While the original formulation can complete videos up to 1K resolution, we further extend it to a progressive scheme for inpainting high-resolution videos such as 2K or 4K videos, as illustrated in Fig. 3. Instead of overfitting the network to the full-resolution frames, we now overfit it using sampled patches. Two problems appear as we increase the resolution. The surrounding pixels may provide limited information as the mask in one specific patch can be especially large. We thus use the upsampled inpainting result from the previous stage as an additional prior. Another issue is the low sampling efficiency, which may result in low training speed. We utilize a boundary-aware sampling (BAS) strategy, which samples more patches from regions around the object boundary. It is based on the fact that in most real-world object removal cases, the pixels around the boundary of the mask contain more valuable information and is of higher importance. More details, including the mask generation, grid-based inference, and BAS, are included in the **Appendix**.

4.2. Video inpainting with a single frame mask

Removing objects in video sequences using only a single frame mask is highly preferred to decrease the involved human labor. We show that by exchanging the input and the output in the above formulation, the proposed scheme can also propagate a given object mask to other frames.

As shown in Fig. 4, the input X' is augmented using contextual information by randomly translating the undesired object on a single given frame X_0. We calculate loss between the prediction mask $\tilde{M}$ and the augmented mask M'_0. Based on the above-mentioned properties, the network f learns to segment similar objects in other frames. Unlike the traditional reference-guided video mask propagation [26, 38], our final goal is not to detect the accurate mask but to remove the desired object. We thus enforce less penalty when we categorize the non-object pixel as object pixel than in the opposite direction as the incorrectly included background pixels can usually be filled in the following inpainting process. We adopt a weighted binary cross-entropy loss $\mathcal{L}_{we}$ which is calculated as $\mathcal{L}_{we}(y, \tilde{y}) = \sum_i y_i \log \tilde{y}_i + \alpha\,(1 - y_i) \log\,(1 - \tilde{y}_i)$. Thus the reconstruction loss is formulated by

$$\mathcal{L}_{rec} = -\,\mathcal{L}_{we}(M'_0 \odot (1 - M_0), \tilde{M} \odot (1 - M_0)), \tag{6}$$

where M'_0 and X'_0 are randomly translated by M_0 and X_0, and α is set to 0.8.

Method	Type	Fix Masks			Object Masks		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
CAP	*external*	28.04	0.906	0.1041	29.37	0.910	0.0483
OPN	*external*	28.72	0.915	0.0872	28.40	0.904	0.0596
STTN	*external*	29.05	0.927	0.0637	29.45	0.918	0.0279
FGVC	*flow-based*	29.68	0.942	0.0564	**33.98**	**0.951**	0.0195
InterVI	*internal*	26.89	0.868	0.1126	27.96	0.875	0.0545
Ours	*internal*	**29.90**	**0.944**	**0.0414**	31.09	0.948	**0.0182**

Table 1: Quantitative comparison on DAVIS.

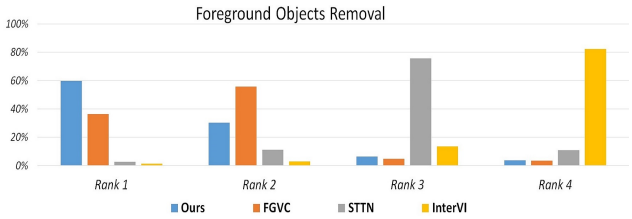

Figure 5: Results of the user study. "Rank x" means the percentage of results from each method being chosen as the x-th best.

5. Experiments

5.1. Training details

For each video sequence in DAVIS [29], we first train the model with only reconstruction loss for about 60,000 iterations and then combine with the regularization loss for another 20,000 iterations for the hard sequences. We use Adam Optimizer with a learning rate of $2e-4$. The training process takes about 4 hours on a single NVIDIA RTX 2080 Ti GPU for a 80-frame video. The training settings of the extension tasks are reported in the **Appendix**.

5.2. Qualitative results

We select five competitive baselines from three categories for fair comparison. CAP [18], OPN [27], and STTN [42] are the most recent external approaches based on deep neural network, FGVC [11] is the newest flow-based optimization method and InterVI [43] is based on deep internal-learning. In Fig. 6, we show the qualitative results on object removal, fixed masks and random object masks. Even if highly-similar content is available in other frames, previous approaches fail to propagate the long-range contextual information and produce blurry and distorted details (e.g., 1st, 3rd examples). They also fail to complete contantly blocked large missing regions and show abrupt artifacts (e.g., 2nd and 4th examples). In contrast, our method can generate realistic textures and reconstruct sharper and clearer structures. Extensive results on other sequences are attached in the **Appendix**.

We also show the mask prediction and inpainting results in Fig. 7. Given a single mask of only one frame, our method can propagate it to other frames of the whole sequence automatically. Note that our internal method usually includes slightly more pixels to ensure that all objects pixels are correctly classified. It performs stably on non-rigid deformation masks and obtains promising results.

5.3. Quantitative results

There are no suitable quantitative metrics for video inpainting due to the ambiguity of ground truth. Nevertheless, we report comparison results of PSNR and SSIM on DAVIS in Table 1, with both fixed and object masks. For the fixed masks, we follow the settings in [39] to simulate a fixed rectangular mask in the center of each frame. The width and height of rectangular masks are both $1/4$ of the original frame. For object masks, we follow the settings in [15, 27, 18] to add imaginary objects on original videos by shuffling DAVIS sequences and masks. In this way, we can simulate test videos with known ground-truth. We show more detailed settings in the **Appendix**. As in Table 1, our method shows substantially comparable performance with state-of-the-art approaches. The flow-based method achieving higher performance on random object mask is understandable since the synthesized object motion is simple which is the ideal case for using the flow-based method as mentioned in [11]. The next section shows the user study on real-world object removal task which demonstrate the advantage that our method is more robust to complex motions.

5.4. User study

Quantitative metrics on synthetic dataset are not sufficient to evaluate the quality of real-world video inpainting. We, therefore, conduct a user study to compare our method with state-of-the-art approaches. Specifically, we randomly select 45 sequences from DAVIS dataset, and slow down the results to 10 FPS for better comparison. We invite 18 participants to rank the results of four methods for each video in terms of visual quality and temporal consistency, and received 810 effective votes totally. Fig. 5 shows that our approach outperforms other methods by a large margin.

6. Ablation study

6.1. Regularization terms

As introduced in Sec. 3, the anti-ambiguity loss aims at recovering the blurry details for the ambiguous cases. Fig. 8 demonstrates examples of moving background, motion blur or complex textures. With the proposed anti-ambiguity term, the background camel's face, the textures of the grass, and the water ripples contain more realistic details.

In Fig. 9, we compare the temporal consistency of the model with and without the stabilization term. The pixels indicated by the yellow line are stacked from all the frames.

Figure 6: Visual comparisons on the DAVIS dataset. Zoom in for details.

Without the proposed stabilization procedure, the temporal graph contains a considerable amount of noise.

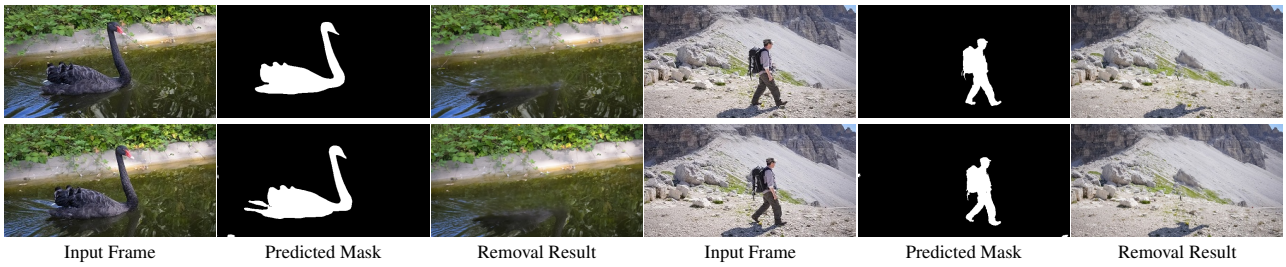

Figure 7: Mask prediction and object removal results given the mask of only the first frame.

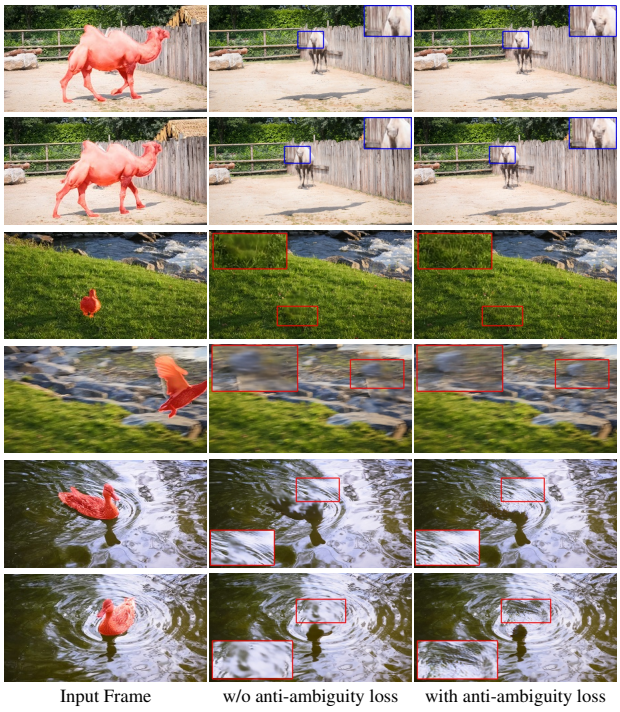

Figure 8: Ablation study on the anti-ambiguity regularization. Zoom in for details.

6.2. Properties analysis

As the first step of our method is to overfit a CNN to memorize all the information in a video sequence, an important aspect is how many parameters are necessary for a specific video sequence which is to analyze Property 2. We use three networks with an increasing number of parameters for testing. With detailed analysis and examples attached in the **Appendix**, we find that using a simple CNN with fewer parameters will suffer a performance drop as the video complexity increase. By involving more parameters to CNN models, the inpainting result on videos with dramatic change or complicated background becomes more realistic.

As mentioned in Sec. 3, we utilize augmented object masks for the internal training as Property 3 requires all the inputs, including the mask shape, to be ideally the same. Another interesting aspect is the performance change if we replace the mask generation strategy with random free-form mask generation. In the new settings, although the training loss diverges, the testing results suffer degradation (in the **Appendix**). This observation shows that using the augmented "object masks", the model can better propagate the information of known regions to unknown regions.

Figure 9: Ablation study on the stabilization regularization.

7. Conclusion

In this work, we propose to formulate video inpainting in a novel way: implicitly propagating long-range information using an overfitted CNN without explicit guidance like optical flow. It succeeds in challenging cases such as large masks, complex motion, and long-term occlusion with the proposed anti-ambiguity and stabilization regularization. We also extend the method to more challenging settings, such as using only a single mask or high-resolution 4K videos. However, there are still two main limitations. Our method is not real-time in training each sequence, like all the other internal learning methods. The other issue is that the details of inpainted areas for the deficiency cases can sometimes be semantically incorrect (See the **Appendix** for failure cases). It is possible to incorporate external semantic information to generate such regions. In the future work, we may explore more on producing better details for deficient frames with inpainted region's confidence.

References

[1] Michael Ashikhmin. Synthesizing natural textures. *SI3D*, 1:217–226, 2001. 2

[2] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Transactions on image processing (TIP)*, 10(8):1200–1211, 2001. 2

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (ToG)*, volume 28, page 24. ACM, 2009. 2

[4] David Bau, Hendrik Strobelt, William Peebles, Bolei Zhou, Jun-Yan Zhu, Antonio Torralba, et al. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 3

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[6] Ya-Liang Chang, Zhe Yu Liu, Kuan-Ying Lee, and Winston Hsu. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9066–9075, 2019. 2

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3213–3223, 2016. 1

[8] Gabriel Eilertsen, Rafal K Mantiuk, and Jonas Unger. Single-frame regularization for temporally stable cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[9] Selim Esedoglu and Jianhong Shen. Digital inpainting based on the mumford–shah–euler image model. *European Journal of Applied Mathematics*, 13(4):353–370, 2002. 2

[10] Yossi Gandelsman, Assaf Shocher, and Michal Irani. double-dip: Unsupervised image decomposition via coupled deep-image-priors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, page 2, 2019. 3

[11] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *European Conference on Computer Vision (ECCV)*, pages 713–729. Springer, 2020. 2, 6, 7

[12] Miguel Granados, Kwang In Kim, James Tompkin, Jan Kautz, and Christian Theobalt. Background inpainting for videos with dynamic objects and a free-moving camera. In *European Conference on Computer Vision (ECCV)*, pages 682–695. Springer, 2012. 2

[13] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2

[14] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017. 2

[15] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5792–5801, 2019. 2, 3, 6

[16] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning (ICML)*, pages 2747–2755. PMLR, 2018. 4

[17] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[18] 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 (ICCV)*, pages 4413–4421, 2019. 2, 3, 6

[19] Chenyang Lei, Yazhou Xing, and Qifeng Chen. Blind video temporal consistency via deep video prior. In *Advances in Neural Information Processing Systems*, 2020. 5

[20] 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 *European Conference on Computer Vision (ECCV)*, pages 728–743. Springer, 2020. 2, 3

[21] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[22] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[23] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 4

[24] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. In *Workshop on the IEEE International Conference on Computer Vision (ICCVW)*, 2019. 2

[25] 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. 2

[26] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9226–9235, 2019. 5

[27] Seoung Wug Oh, Sungho Lee, Joon-Young Lee, and Seon Joo Kim. Onion-peel networks for deep video completion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4403–4412, 2019. 2, 3, 6, 7

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[29] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016. 1, 6

[30] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[31] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[32] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani. Ingan: Capturing and remapping the "dna" of a natural image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[33] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[35] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang. Video inpainting by jointly learning temporal structure and spatial details. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 5232–5239, 2019. 2

[36] Tengfei Wang, Hao Ouyang, and Qifeng Chen. Image inpainting with external-internal learning and monochromic bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5120–5129, 2021. 2, 3

[37] Yonatan Wexler, Eli Shechtman, and Michal Irani. Space-time video completion. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2004. 2

[38] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7376–7385, 2018. 5

[39] 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 (CVPR)*, pages 3723–3732, 2019. 2, 3, 6

[40] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[42] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision (ECCV)*, pages 528–543. Springer, 2020. 2, 3, 6, 7

[43] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2720–2729, 2019. 2, 3, 6, 7

[44] Ruisong Zhang, Weize Quan, Baoyuan Wu, Zhifeng Li, and Dong-Ming Yan. Pixel-wise dense detector for image inpainting. In *Computer Graphics Forum*, 2020. 2