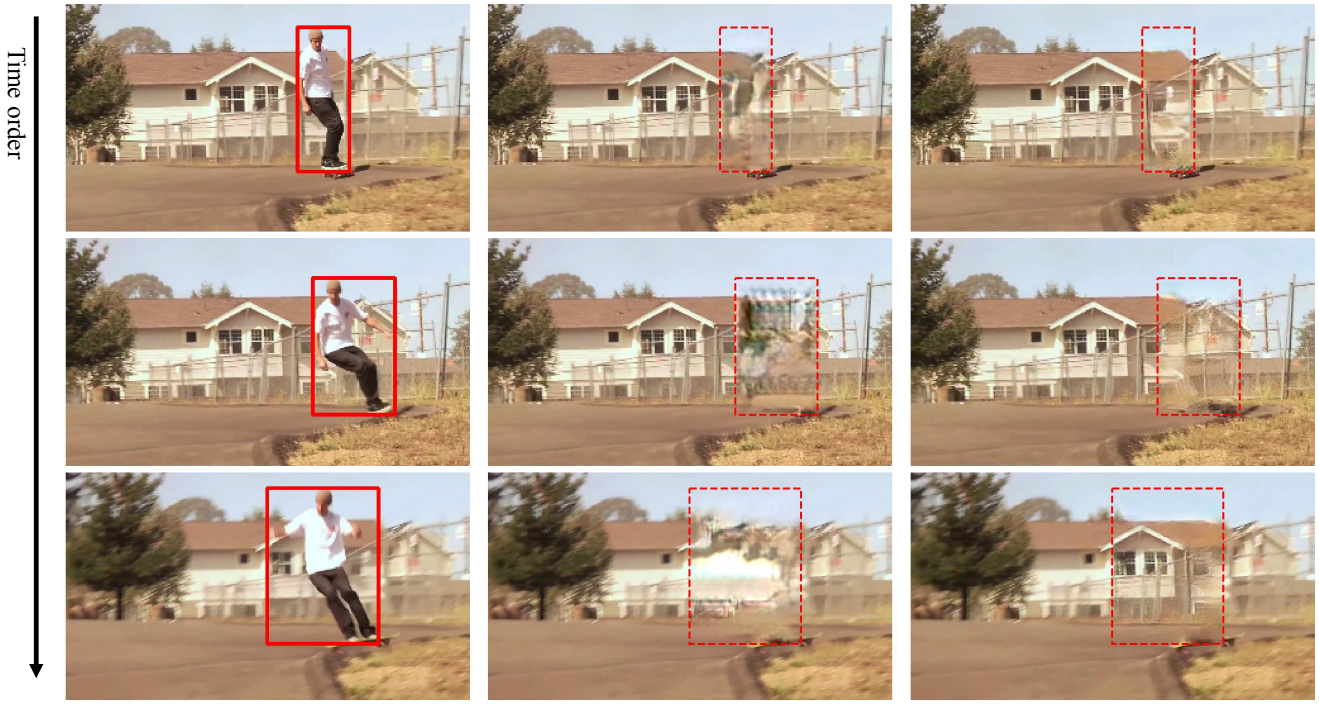
VORNet: Spatio-temporally Consistent Video Inpainting for Object Removal

Ya-Liang Chang Zhe Yu Liu Winston Hsu

National Taiwan University, Taipei, Taiwan
{yaliangchang, zhe2325138}@cmlab.csie.ntu.edu.tw, whsu@ntu.edu.tw



(a) Input video (b) Image-based (c) Ours

Figure 1: Video object removal results. (a) The input video and its foreground bounding boxes to remove, marked in red. (b) State-of-the-art image-based inpainting model by Yu *et al.* [50]. (c) Our results. Applying image-based algorithms on the video inpainting task often leads to temporal inconsistency, where the content is different in each frame, *e.g.,* the windows are missing in the last frame of (b). Our deep learning based architecture could improve both the spatial and temporal consistency of image-based inpainting models. Best viewed with color and zoom-in. See http://bit.ly/2GkW9Kr for the video.

Abstract

Video object removal is a challenging task in video processing that often requires massive human efforts. Given the mask of the foreground object in each frame, the goal is to complete (inpaint) the object region and generate a video without the target object. While recently deep learning based methods have achieved great success on the image inpainting task, they often lead to inconsistent results between frames when applied to videos. In this work, we propose a novel learning-based Video Object Removal Network (VORNet) to solve the video object removal task in a spatio-temporally consistent manner, by combining the optical flow warping and image-based inpainting model. Experiments are done on our Synthesized Video Object Removal (SVOR) dataset based on the YouTube-VOS video segmentation dataset, and both the objective and subjective evaluation demonstrate that our VORNet generates more spatially and temporally consistent videos compared with existing methods.

1. Introduction

Removing undesired objects in videos is crucial to many applications, such as movie post-production and video editing. While manually removing objects in a video requires

substantial human efforts, automatic video object removal could save a great amount of time. Given the region of the foreground object in each frame, the goal of automatic video object removal is to fill in, or inpaint, the foreground region with background content and generate a video without the target object. Automatic video object removal is a very challenging task since it requires both spatial and temporal consistency; the inpainted region must fit in the background seamlessly in diverse scenes, and it should remain consistent appearance in the following frames where its surroundings may change significantly. Some examples of inconsistent frames include flickering and distortion (see Fig. 1(b) and Fig. 5(b)(c)).

Video object removal could be viewed as an extension of the image/video inpainting task. Early patch-based inpainting methods [10, 11, 39] divide images into small patches and recover the masked region by pasting the most similar patch somewhere in the image/video. These methods could generate authentic results but they are usually very time-consuming due to the complexity of neighbor-finding algorithms [26]. In addition, patch-based methods assume there is a reference for the missing part and often fail to recover non-repetitive and complex region (e.g, they cannot recover a missing face well [28]).

On the other hand, deep learning based image inpainting models could estimate the missing parts based on the training data and generate novel results with impressive quality [29, 44, 45, 48, 50]. One naive idea is to solve the video object removal problem by applying these image inpainting models to each frame to recover the foreground region. Nonetheless, when applied to videos, these image-based methods would generate temporally inconsistent results that cause flickering or distorted videos, since they do not consider the temporal relation between frames and treat them independently.

We propose a novel learning-based architecture for video object removal that could take advantage of existing state-of-the-art image-based inpainting models and generate visually plausible frames in a temporally consistent manner. The core idea is to combine the information from previous frames and generated result in current frame. For previous information, we use the optical flow to capture background motion and recover removed foreground part by warping the previous background accordingly. For the constantly occluded region, existing image-based inpainting models could generate plausible results. Based on these candidates, we design a refinement network to select and refine them to derive a spatially and temporally consistent result.

Since there is no existing dataset for video object removal, we build a large-scale Synthesized Video Object Removal (SVOR) dataset based on the YouTube-VOS [43] video segmentation dataset. A variety of foreground segmentation and background videos are selected from YouTube-VOS videos and synthesized to 1958 video-with-target and video-without-target pairs. We train our VORNet on the SVOR dataset with reconstruction loss, perceptual loss and two designed GAN losses and evaluate the quality of videos with mean square error, SSIM [37], a learned perceptual metric [51] and visual results. We show that the proposed method could improve the perceptual quality and temporal stability. Our VORNet processes frames online, sequentially, does not require post-processing and could deal with videos in various lengths.

Our contributions could be summarized with the following points:

- We propose a novel Video Object Removal Network (VORNet) to remove undesired objects in videos. To our knowledge, VORNet is the first learning based model for video object removal. It could generate visually plausible and temporally coherent result online, without post-processing.
- We design a combination of spatial content losses and temporal coherent loss based on GAN structure to train our model, which could improve the spatio-temporal quality of generated videos.
- We create the first large-scale Synthesized Video Object Removal (SVOR) dataset based on the YouTube-VOS dataset. The SVOR dataset contains a huge variety of motions and scenes that could be used for training and evaluation in further research. The dataset is publicly available here: `http://bit.ly/2P3n2oH`.

2. Related Work

Image Inpainting Image inpainting was first introduced in [3] as a general image processing problem that aims to recover the damaged or missing region of an image. Subsequently, a great amount of research is done for image inpainting [16] with diffusion-based [2, 3] and patch-based [5, 10, 12, 38] algorithms. These traditional methods perform well on simple structure but are very limited to complex objects, large missing area and non-repetitive texture where similar reference may not exist.

In recent years, learning-based models demonstrate promising results with the help of deep convolutional neural network (CNNs). These models learn image features in the training data and are thus capable of generating realistic content that may not exist in the unmasked area, such as faces [28, 50], complex objects [31] and natural scenes [19, 50]. Xie *et al.* [40] is the first to train convolutional neural networks for image denoising and inpainting on small regions. Pathak *et al.* [33] further extend the work to a larger region by an encoder-decoder structure. Also, to improve blurry effect caused by the l_2 loss, Pathak *et al.* [33] introduce the idea of adversarial loss from the generative adversarial network (GAN) [14] where a generator that aims to create

real images to fool the discriminator and a discriminator that strikes to tell the fidelity of generated images are jointly trained.

More recently, Yu *et al.* [50] add a contextual attention layer to and several improvements on network design to produce higher-quality images. It is trained on the diverse Places2 [53] dataset and achieve state-of-the-art result, so we shall take it as our inpainting network and a baseline.

Yan *et al.* [44], Yu *et al.* [49] and Lui *et al.* [29] also manage to solve the problem of inpainting irregular holes. However, since precise segmentation of an object in a video may not be derived easily, we focus only on inpainting bounding box region of the object in this work. We assume the foreground bounding boxes are given as they could be easily derived by object tracking methods or human annotations.

Video Inpainting Video inpainting is generally viewed as an extension of the image inpainting task with larger search space and temporally consistent constraints. Early works [15, 38, 32] are mainly extensions of patch-based methods from image inpainting, where images are split into small patches and the masked region is recovered by pasting the most similar patch somewhere in the image/video. Wexler *et al.* [38] consider the video inpainting task as a global optimization problem that all missing portion could be filled in with patches from the available parts of the video with enforced global spatio-temporal consistency. Wexler *et al.* [38] propose an iterative approach to solve the global optimization problem and yield magnificent results in an automatic way. However, due to the large search space and the complexity of the nearest neighbor search algorithm, their method is extremely slow that processing a few seconds of video may take days to compute. Also, the assumption that there exists a similar patch that could fill in the missing region may not hold under circumstances like a long-lasting occlusion, a moving camera or masked regions with semantic ambiguity. [20].

The following works try to solve these issues. Newson *et al.* [32] extend the work of Wexler *et al.* [38] by accelerating the algorithm, adding texture features and initialization scheme. Ebdelli *et al.* [13] also limit the search space in an aligned group of frames to reduce computational time. Huang *et al.* [17] address the moving camera problem by estimating the optical flow and color in the missing regions jointly. However, the computation time of these methods is still longer than per-frame processing after acceleration. In addition, patch based models still lack modeling distribution of real images, so they fail to recover unseen parts in the video. Our data-driven method could solve both issues by learning the distribution of frames and generate realistic videos by forward inference, without searching.

Video Temporal Consistency Video temporal consistency aims to solve the flickering problem when applying different kinds of image-based models like photo enhancement [8] colorization [52], style transfer [18, 30] and general image-to-image translations [22, 54] to videos. Generally, it could be divided into task-independent and task-specific methods.

Task-independent approaches [4, 25, 46] aim to use a single model to handle multiple applications with the video temporal consistency problem. Among them, the recent work by Lai *et al.* [25] propose an efficient method using a deep network that could generate impressive temporally coherent videos in real time, given various types of temporally inconsistent inputs and their original unprocessed videos as reference. They use the FlowNet2 [21, 34] to estimate their temporal loss to train the model. However, for the video object removal task, the method of Lai *et al.* [25] does not work because the inpainted region in the unprocessed video is occupied by the foreground object, which not be used as a reference for temporal loss. Instead, our refinement network utilizes the warping network and temporal discriminator to generate temporal consistent results.

Task-specific approaches like [7, 27, 47] develop different strategies according to each domain. Some attempt to design specific temporal filters [1] or embed optical flow estimation to capture information of motion [7]. Recently, Xie *et al.* [41] design a temporal discriminator aside from a normal spatial one for the fluid flow super-resolution task. It utilizes motion from low-resolution video to generate temporally consistent high-resolution fluid flow video, but there is no such reference in the video inpainting task. Alternatively, we extend this work to design our temporal discriminator without reference for video inpainting.

Wang *et al.* [36], concurrent with our work, propose a deep learning architecture to address the inconsistent problem in video inpainting. The method uses a 3D convolutional network to learn the temporal relation and generate coarse temporally consistent images for the masked area, and refine them with a 2D convolutional network. Although results of Wang *et al.* [36] are temporally consistent, their model could not generate clear videos for a diverse dataset as only the L_1 loss is used for training. Our VORNet could utilize existing image-based inpainting models and improve the video quality by the combinations of different loss functions .

3. Video Object Removal

Our VORNet takes as input the video-with-target frames $\{I_t \mid t = 1 \dots n\}$ and the target bounding box mask in each frame $\{M_t \mid t = 1 \dots n\}$ in sequence and generate the output video-without-target frames $\{O_t \mid t = 1 \dots n\}$. The model is composed of three parts: the warping network, the inpainting network, and the refinement network (see Fig. 2). The core concept is to use the information from other frame (warping network) and generated frame (inpainting network), combine and refine them in a spatio-temporally

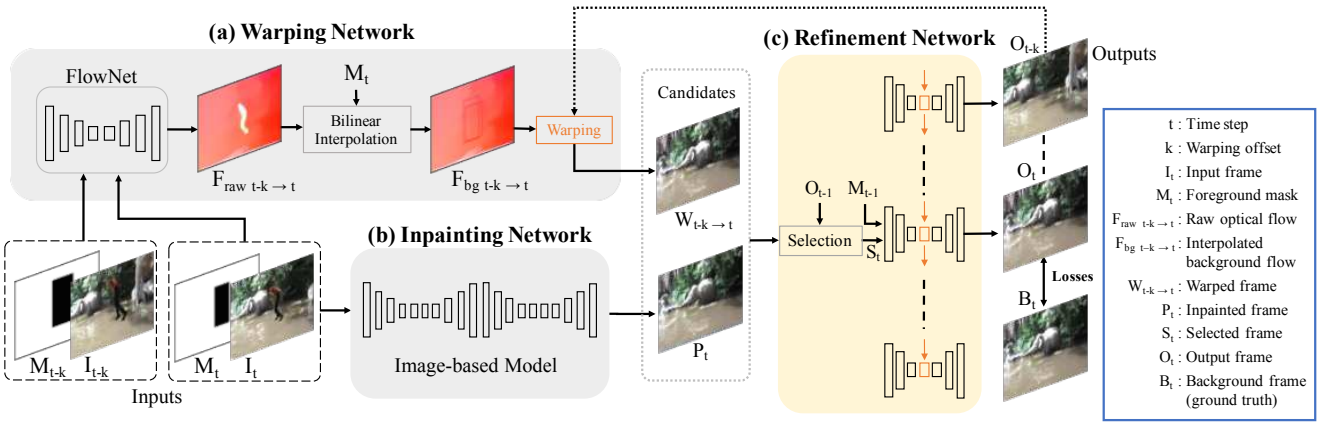


Figure 2: Our VORNet architecture and notations. (a) The warping network aims to collect information from other frames (see details in Fig. 3). (b) The inpainting network intends to estimate the missing parts, by using the generative model to create a possible image according to its surroundings. (c) The refinement network is designed to combine the information from other frames and the estimated frame, by selecting and refining candidates (see details and losses in Fig. 4). The model runs in a recurrent way; the output frame is used to create warping candidates of the next frame, and the current state in refinement network would propagate to the next frame. Best viewed with color.

coherent way (refinement network).

3.1. The Warping Network

The warping network aims to collect information from other frames. For example, if the target object to be removed is static in two consecutive frames and the background is moving rightward, we could know that the foreground region of the second frame should be filled in with the background in its left side in the first frame (see Fig. 3).

To estimate these relative motions between two input frames I_{t-k} and I_t, we use FlowNet2 [21, 34] pre-trained on the MPI-Sintel Dataset [6] to calculate the raw optical flow $F_{raw_{t-k \to t}}$ between them.

However, for $F_{raw_{t-k \to t}}$, the foreground region is derived from the pasted foreground object, which could not represent the background motion between the last frame and this frame. To address this issue, we remove the foreground region in the raw optical flow and apply simple bilinear interpolation to fill in the removed region and recover the background optical flow $F_{bg_{t-k \to t}}$. We do not adopt the learning based method in this component because the performance is not as expected considering its cost.

Finally, we warp O_{t-k} to the M_t region with inpainted flow $F_{bg_{t-k \to t}}$ using the warping operation as [35] and send it to the refinement network as a candidate. We have candidates with different k so that we could get information from temporally closer and further neighbors.

3.2. The Inpainting Network

The inpainting network intends to estimate the missing background. It could be any model that recover the masked part of input videos, including learning based and patch based ones. To estimate occluded regions that patch-based models could not handle, we adopt the generative inpainting network from Yu *et al.* [50] pre-trained on the Places2 dataset [53] and fine-tune on our SVOR dataset. It consists of a coarse network that generates a coarse result from the masked input image and a refinement network that turns the coarse result to the final output with contextual attention. Details for the model could be found in the supplementary material.

3.3. The Refinement Network

The refinement network is designed to combine the candidates from warping network and inpainting network. Given candidates from warped frames and the inpainted frame, the refinement network will select the top 1 candidate S_t to generate the final output frame O_t with the mask M_t. To maintain temporal consistency, the selection is done by choosing the candidate that is closest to the previous result in the feature level (LPIPS [51] distance, see Sec 4.3). Finally, losses are computed using the output O_t and the background frame B_t.

As shown in Fig. 4, the refinement network includes three convolutional layers that encode the candidate frames and mask, a convolutional LSTM [42] layer that propagates temporal features, and three transposed convolutional layers to reconstruct the image. Skip connections are added between convolutional layers and corresponding transposed convolutional layers.

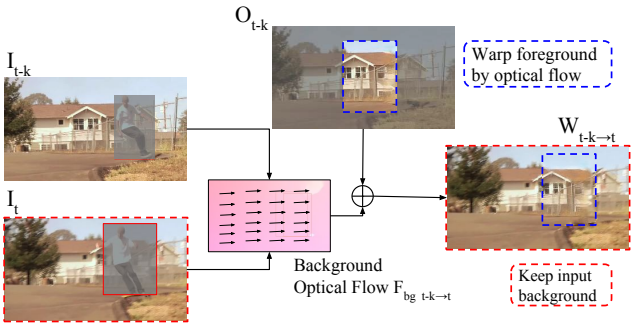


Figure 3: Concept of the warping network. In the current input frame I_t, the window is hard to be reconstructed for image-based models since it is occluded by the man, but we could easily know that the window should be there seeing the previous frame I_{t-k}. Based on this idea, the warping network estimates the motion (optical flow) between the I_t and its kth previous frame I_{t-k}, and generate the warped frame $W_{t-k\to t}$ by warping the corresponding foreground area in the previous output O_{t-k} (marked in blue) to I_t. Note that we use O_{t-k} to warp the region instead of I_{t-k} because I_{t-k} may include the foreground. Best viewed with color and zoom-in.

3.4. Loss Functions

Our aim is to recover the background region which is masked by the foreground object. This is a challenging task because we need to consider both the spatial and temporal consistency. Accordingly We propose to train our VORNet with spatial content loss from low-level to high-level and temporally coherent loss.

Spatially discounted reconstruction loss. l_1 loss focuses on the lowest level of pixel difference. We embrace the spatially discounted reconstruction loss in [50]

$$L_{l_1} = \mathbb{E}_{t,x,y}[|O_{t_{x,y}} - B_{t_{x,y}}|(\gamma_{t_{x,y}})^d] \tag{1}$$

where x and y denote the pixel indexes of a frame, and the loss in each pixel is weighted by γ^d according its distance d to the nearest boundary. It is more suitable for image inpainting task compared to naive l_1 loss since pixels closer to the boundary should match the background, while the middle part could have more diversity.

VGG perceptual loss. One problem about l_1 loss in a generative task is that it usually produces blurry results, because it is hard for the model to minimize the l_1 loss when generating a sharp and vivid image.

Therefore, we adopt a VGG-net pre-trained on a classification task [23] to compute the perceptual distance between generated and ground truth images as one of our spatial loss

$$L^{\phi,j}_{perc} = \mathbb{E}_t[\|\phi_j(O'_t) - \phi_j(B'_t)\|^2_2] \tag{2}$$

where ϕ and j denote the VGG network and its layer index respectively. The O'_t and B'_t are the output and background image cropped to the masked area. The perceptual loss emphasizes on the higher level of difference like style or textures instead of pixels.

PatchGAN loss. To motivate our model to generate realistic images, we use the PatchGAN discriminator [22] as our spatial discriminator D_s, while the refinement model could be viewed as a generator G. The Patch GAN loss is defined as

$$\begin{aligned} L_{GAN_s}(G, D_s) = \mathbb{E}_t[log(D_s(B'_t))] \\ + \mathbb{E}_t[log(1 - D_s(G(S'_t)))] \end{aligned} \tag{3}$$

where S'_t and B'_t denote the selected and background image cropped to the masked area. While the l_1 loss focus on low-frequency structure, PatchGAN discriminator penalizes local patches only for the high-frequency structure [22].

Temporal GAN loss. The above losses are for the image quality only, while we need a temporal constraint to generate content coherent videos. To solve this problem, we design a temporal discriminator to train our model. A similar idea could be seen in a recent fluid flow super-resolution work [41], which propose the TempoGAN to generate temporally coherent high-resolution fluid flow video utilizing flow motion in low-resolution one. While for video inpainting there is no low-resolution reference, our temporal discriminator estimates the consistency score by the differences of consecutive frames in the feature level. It takes the features from output frames and the foreground masks as inputs, calculates siamese features [9] differences, further extract features and estimate the final consistent score (see Fig. 4b1). With the proposed temporal discriminator D_t, the temporal GAN loss is defined as:

$$\begin{aligned} L_{GAN_t}(G, D_t) = \mathbb{E}_t[log(D_t(B_t))] \\ + \mathbb{E}_t[log(1 - D_t(G(S_t)))] \end{aligned} \tag{4}$$

Overall loss. The overall loss function to train our VORNet is defined as:

$$\begin{aligned} L = \lambda_{l_1} \times L_{l_1} + \lambda_{perc} \times L^{\phi,j}_{perc} \\ + \lambda_{G_s} \times L_{G_s} + \lambda_{G_t} \times L_{G_t} \end{aligned} \tag{5}$$

where λ_{l_1}, λ_{perc}, λ_{G_s} and λ_{G_t} are the weights for reconstruction loss, perceptual loss, spatial GAN loss and temporal GAN loss, respectively.

4. Experimental Results

4.1. Dataset

We build our Synthesized Videos for Object Removal (SVOR) dataset based on the YouTube-VOS dataset [43],

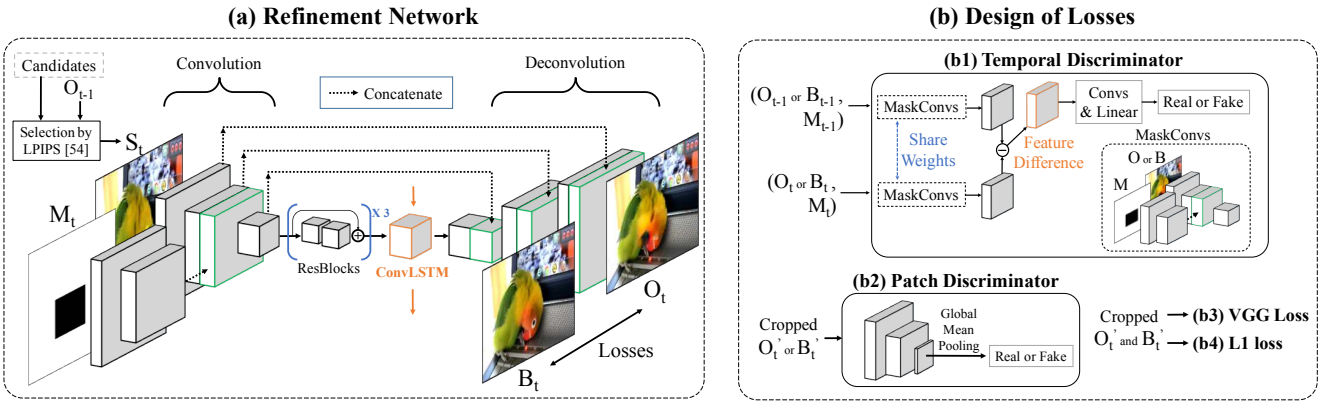


Figure 4: Refinement network architecture and losses design. (a) The network select the S_t from candidates (warped and inpainted frames) by the closest LIPIS [51] distance to the last output, and use S_t and the foreground mask M_t to generate the final output O_t. The temporal information is propagated by the convolutional LSTM (ConvLSTM) [42] in high-level features. Lastly, losses between the output O_t and the ground truth background B_t will be computed. (b1) The temporal discriminator will estimate the real/fake of output or ground truth frames by the siamese features [9] differences with the last frame. (b2) PatchGAN focus only on masked area. Best viewed with color and zoom-in.

which is a large-scale dataset for video segmentation including a huge variety of moving objects, camera view and motion types. The YouTube-VOS dataset consists of 4,453 videos and 7,822 unique objects including humans with diverse activities, animals, vehicles, accessories and some common objects. Each video is about 3 to 5 seconds, and up to five human annotated object segmentation masks are given every five frames in a 30 FPS frame rate.

Since it is not likely to get the real ground truth for the video object removal task , we utilize segmentation in the YouTube-VOS training set to synthesize training video-with-target/video-without-target pairs. After manually filter out videos where the annotated object is only partially in the screen, occupied more than one-half of the screen or smaller than 30×30 pixels, 1,958 videos are used to synthesize our input videos.

We split these videos into 1,858 training and 100 testing videos, and create 1,858 and 100 pairs among them (there could be significantly more pairs if existing duplicated foreground/background videos). Each synthesized video pair is composed of one foreground video and one background video from the YouTube-VOS dataset. We take the first object segmentation mask in the foreground video as the target objects and paste it to the background video. Consequently, it becomes the input synthesized video-with-target, and the background video is viewed as the ground truth video-without-target.

In the SVOR dataset, the foreground object may be static or moving, and its size could vary in a single video. Also, the background may be shaky, following an object or changing the brightness. Some background objects could also be originally in the foreground region or moving toward the region, so the SVOR dataset is very diverse and challenging.

4.2. Implementation Details

Our model is implemented with Pytorch 0.4.1 and trained on our SVOR dataset in 320×180, at most 15 frames for each video. Warping temporal distance ks are set to be $\{1, 3, 5\}$. The $relu3_3$ layer is used for the VGG loss. γ for L_{l_1} is set to be 0.99. The patch size for PatchGAN is 15×15. Loss weights λ_{l_1}, λ_{perc} and λ_{G_s} are set to be 1. λ_{G_t} is 0.01. Other details could be found in the supplementary material.

4.3. Quantitative and Qualitative Comparisons

We compare the proposed method with the well-known patch-based video inpainting methods [32, 17] with patch size $3 \times 3 \times 3$, the state-of-the-art image-based inpainting model [50] pre-trained on the Places2 dataset and fine-tuned on our SVOR dataset and the two-stage learning based video inpainting model [36]. In general, our VORNet performs better than the four benchmarks quantitatively and qualitatively.

We report the evaluation in terms of mean square error (MSE) and structural similarity (SSIM) [37], which are commonly used in inpainting tasks [29, 44, 50]. However, these traditional evaluation metrics may not represent the perceptual distance well (i.e., they prefer blurry images than partially shifted, distorted images). As a result, we also use the recently proposed Learned Perceptual Image Patch Similarity (LPIPS) [51] to estimate perceptual distance. LPIPS calibrates features of ImageNet classification networks and corresponds more to human perception. We take the model calibrated on the AlextNet [24] as suggested [51]. The quan-

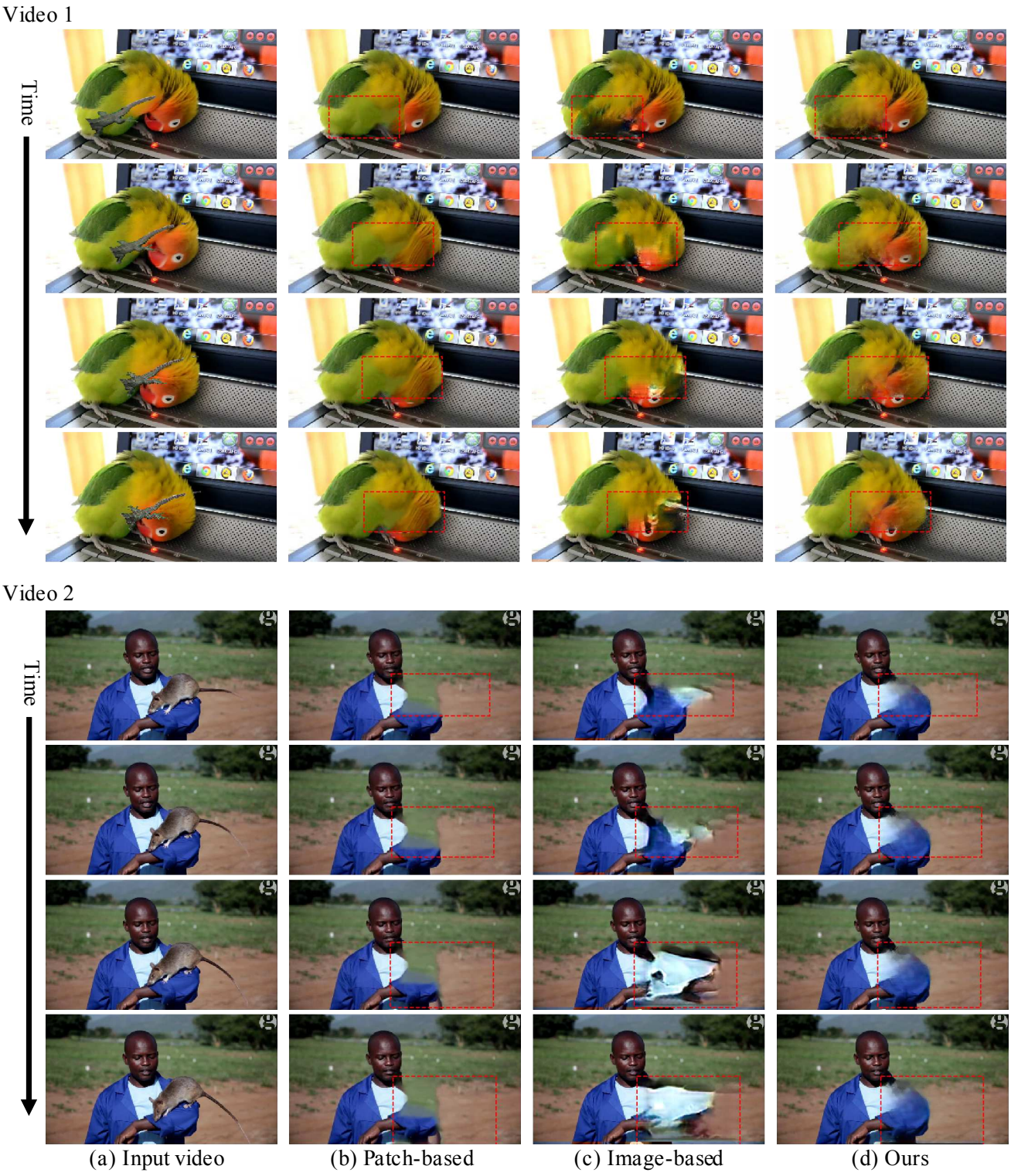


Figure 5: **Visual results** compared with (b) state-of-the-art patch-based video inpainting by Huang *et al.* [17] and (c) state-of-the-art image-based inpainting by Yu *et al.* [50]. **Video 1**: the first frame is spatially consistent for all methods. However, for (b) and (c), the bird lose its eye, while ours keeps it intact by the warping network. **Video 2**: the man's shoulder is filled with grass for the patch-based method (b) as it could not tell the surroundings. For image-based method (c), we can see that the second and third frame is very different, while our results remain temporally consistent. Best viewed in color and zoom-in.

titative result of 100 synthesized video in the testing set could be seen in Table. 1. We could see that our model outperforms the four benchmarks.

Since the quantitative result of frames may not represent the temporal consistency, we also evaluate qualitatively on 100 testing videos. The visual comparison could be seen in Fig. 5. Our results remain spatio-temporally stable as surroundings change, while results of other methods become distorted or inconsistent. More visual comparisons with all baselines [32, 17, 50, 36] could be found in https://bit.ly/2I7WbID (synthesized), https://bit.ly/2GdnbUX (real) and the supplementary material.

4.4. Ablation Study

To evaluate the contribution of each component in the proposed model, we conduct ablation study on main components including the warping network, VGG loss, spatial discriminator and temporal discriminator. The result is shown in Table. 2. We could see that the warping network play an important role in our VORNet, while each loss has some effects on the result. Specifically, if VGG loss is removed, the model would generate sharp images disregarding the surrounding content; if the spatial GAN loss is removed, there would be some unnatural repetitive patterns that could reduce MSE; if the temporal GAN loss is removed, the result would be slightly temporally inconsistent. Corresponding visual comparisons could be found in the supplementary material.

4.5. Execution Time

The execution time is evaluated on a machine with a Intel Xeon E5-2650 v3 CPU (128G RAM) and two Nvidia Tesla K80 GPUs. The speed of VORNet is 2.5 frame per second (FPS), slower than the Yu *et al.* [50] (11 FPS) due to FlowNet2 [21, 34] full-model optical flow estimation (ours is 7 FPS with FlowNet2-S [21, 34]), while faster than the video inpainting method [32] (0.15 FPS) with patch size $3 \times 3 \times 3$ since it runs on the CPU. Note that our VORNet does not require post-processing and can run online, without peeking the future frames.

4.6. Limitations and Discussion

Our model relies on the optical flow to get information from the previous frames, which results in extra execution time and parameters. In addition, the state-of-the-art networks for optical flow inference still could not capture object motions in detail and there is unavoidable occlusion problem, which make the warped frames blurry. A possible solution is to design a temporal attention and warping network that could replace the optical flow warping. The model could be trained in an end-to-end way and the performance may be improved. Still, the proposed method is the first learning-based architecture for the video object removal task and produces state-of-the art results.

Method	MSE ↓	SSIM ↑	LPIPS ↓
Huang *et al.* [17]	0.01665	0.6967	0.2385
Newson *et al.* [32]	0.02152	0.6577	0.2409
Yu *et al.* [50]	0.02009	0.6896	0.2249
Wang *et al.* [36]	0.01566	0.6749	0.3915
VORNet (Ours)	**0.01560**	**0.7260**	**0.1889**

Table 1: Quantitative results of the proposed network, state-of-the-art patch-based video inpainting [17, 32], image inpainting [50] and learning-based video inpainting [36] methods. We could use original background videos as ground truth to calculates these metrics since we evaluate on our synthesized dataset.

Warping network	VGG loss	Spat. Disc.	Temp. Disc.	MSE ↓	LPIPS↓
	✓	✓	✓	0.01807	0.2576
✓		✓	✓	0.01846	0.2460
✓	✓		✓	**0.01314**	0.2291
✓	✓	✓		0.01669	0.2039
✓	✓	✓	✓	0.01560	**0.1889**

Table 2: Ablation study of the components including warping network, VGG loss, spatial discriminator and temporal discriminator. MSE and LPIPS [51] distance with the ground truth is calculated for the 100 testing pairs.

5. Conclusion

In this work, we propose a novel Video Object Removal Network (VORNet) for the video object removal task, utilizing existing image-based inpainting model and enhance the spatial and temporal consistency. To our knowledge, our VORNet is the first to introduce learning-based method to the video object removal task. We design spatial and temporal GAN losses and train the proposed model on our Synthesized Video Object Removal Dataset (SVOR) based on the YouTube-VOS video segmentation dataset. Our model is learning based, runs online, faster than patch-based video inpainting method and does not require post-processing. Evaluation on perceptual distance, visual result and user studies show that our model achieves state-of-the-art results compared to existing methods.

Acknowledgement

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2634-F-002-004. We also benefit from the NVIDIA grants and the DGX-1 AI Supercomputer. We are grateful to the National Center for High-performance Computing.

References

[1] T. O. Aydin, N. Stefanoski, S. Croci, M. Gross, and A. Smolic. Temporally coherent local tone mapping of hdr video. *ACM Transactions on Graphics (TOG)*, 33(6):196, 2014. 3

[2] M. Bertalmio, A. L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001. 2

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[4] N. Bonneel, J. Tompkin, K. Sunkavalli, D. Sun, S. Paris, and H. Pfister. Blind video temporal consistency. *ACM Transactions on Graphics (TOG)*, 34(6):196, 2015. 3

[5] R. Bornard, E. Lecan, L. Laborelli, and J.-H. Chenot. Missing data correction in still images and image sequences. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 355–361. ACM, 2002. 2

[6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 4

[7] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua. Coherent online video style transfer. In *Proc. Intl. Conf. Computer Vision (ICCV)*, 2017. 3

[8] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018. 3

[9] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005. 5, 6

[10] A. Criminisi, P. Perez, and K. Toyama. Object removal by exemplar-based inpainting. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2003. 2

[11] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004. 2

[12] I. Drori, D. Cohen-Or, and H. Yeshurun. Fragment-based image completion. In *ACM Transactions on graphics (TOG)*, volume 22, pages 303–312. ACM, 2003. 2

[13] M. Ebdelli, O. Le Meur, and C. Guillemot. Video inpainting with short-term windows: application to object removal and error concealment. *IEEE Transactions on Image Processing*, 24(10):3034–3047, 2015. 3

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[15] M. Granados, J. Tompkin, K. Kim, O. Grau, J. Kautz, and C. Theobalt. How not to be seenobject removal from videos of crowded scenes. In *Computer Graphics Forum*, volume 31, pages 219–228. Wiley Online Library, 2012. 3

[16] C. Guillemot and O. Le Meur. Image inpainting: Overview and recent advances. *IEEE signal processing magazine*, 31(1):127–144, 2014. 2

[17] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (TOG)*, 35(6):196, 2016. 3, 6, 7, 8

[18] X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 3

[19] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017. 2

[20] S. Ilan and A. Shamir. A survey on data-driven video completion. In *Computer Graphics Forum*, volume 34, pages 60–85. Wiley Online Library, 2015. 3

[21] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, page 6, 2017. 3, 4, 8

[22] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 3, 5

[23] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 6

[25] W.-S. Lai, J.-B. Huang, O. Wang, E. Shechtman, E. Yumer, and M.-H. Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision*, 2018. 3

[26] T. Le, A. Almansa, Y. Gousseau, and S. Masnou. Motion-consistent video inpainting. In *ICIP 2017: IEEE International Conference on Image Processing*, 2017. 2

[27] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In *ACM transactions on graphics (tog)*, volume 23, pages 689–694. ACM, 2004. 3

[28] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 2

[29] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. *arXiv preprint arXiv:1804.07723*, 2018. 2, 3, 6

[30] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. *CoRR, abs/1703.07511*, 2, 2017. 3

[31] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2

[32] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez. Video inpainting of complex scenes. *SIAM Journal on Imaging Sciences*, 7(4):1993–2019, 2014. 3, 6, 8

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2

[34] F. Reda, R. Pottorff, J. Barker, and B. Catanzaro. flownet2-pytorch: Pytorch implementation of flownet 2.0: Evolution of optical flow estimation with deep networks. https://github.com/NVIDIA/flownet2-pytorch, 2017. 3, 4, 8

[35] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 4

[36] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. *arXiv preprint arXiv:1806.08482*, 2018. 3, 6, 8

[37] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 2, 6

[38] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):463–476, 2007. 2, 3

[39] J. Wu and Q. Ruan. Object removal by cross isophotes exemplar-based inpainting. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 810–813. IEEE, 2006. 2

[40] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in neural information processing systems*, pages 341–349, 2012. 2

[41] Y. Xie, E. Franz, M. Chu, and N. Thuerey. tempogan: A temporally coherent, volumetric gan for super-resolution fluid flow. *ACM Transactions on Graphics (TOG)*, 37(4):95, 2018. 3, 5

[42] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 4, 6

[43] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2, 5

[44] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan. Shift-net: Image inpainting via deep feature rearrangement. *arXiv preprint arXiv:1801.09392*, 2018. 2, 3, 6

[45] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 2

[46] C.-H. Yao, C.-Y. Chang, and S.-Y. Chien. Occlusion-aware video temporal consistency. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 777–785. ACM, 2017. 3

[47] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez. Intrinsic video and applications. *ACM Transactions on Graphics (TOG)*, 33(4):80, 2014. 3

[48] R. A. Yeh, C. Chen, T.-Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *CVPR*, volume 2, page 4, 2017. 2

[49] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018. 3

[50] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. *arXiv preprint*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint*, 2018. 2, 4, 6, 8

[52] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*, 2017. 3

[53] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3, 4

[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017. 3