Supplementary Material:
Onion-Peel Networks for Deep Video Completion

Seoung Wug Oh
Yonsei University

Sungho Lee
Yonsei University

Joon-Young Lee
Adobe Research

Seon Joo Kim
Yonsei University

1. Temporal Consistency Networks

For the video completion, we post-process frame-by-frame results with the temporal consistency network (TCN) [5]. We modified the original work [5] to match our purpose: stabilizing the inpainted videos. Here, we provide more detailed descriptions of the post-processing networks.

Network Design The network structure of TCN is shown in Fig. 1. The networks consists of the encoder, the convolutional GRU [2], and the decoder. The encoder inputs are the previous stabilized frame (P_{t-1}), the current frame to be stabilized (I_t), and their object masks. The output is the stabilized current frame (P_t). The convolutional GRU [2] is employed to capture a long-term temporal consistency. Skip-connections links the encoder and the decoder features. All the convolutional layer is the gated convolutional layer [10].

Loss Function. The networks is trained to balance between the temporal stability with the previous frame and the perceptual similarity with the current frame.

The temporal stability loss is defined as a pixel distance toward warped previous output:

$$\mathcal{L}_{\text{ts}} = \|M \odot (P_t - \hat{P}_{t-1})\|_1, \tag{1}$$

where $\hat{P}_{t-1}$ is the previous output P_{t-1} warped by the optical flow $F_{t-1 \Rightarrow t}$ and M is a visibility map. The optical flow is computed from the ground truth frames Y_t, Y_{t-1} (training frames without holes). We used PWC-Net for computing the optical flow [8]. The visibility map M is defined as $M = \exp(-100\|Y_t - \hat{Y}_{t-1})\|_2^2)$.

The perceptual similarity loss is defined as follows:

$$\mathcal{L}_{\text{ps}} = \|\phi_5(I_t) - \phi_5(P_t)\|_1, \tag{2}$$

where $\phi_s(\cdot)$ is the mapping to s-th pooled feature map of VGG-16 network [7] pre-trained on ImageNet.

The total loss is the weighted summation of two:

$$\mathcal{L}_{\text{total}} = 15 \cdot \mathcal{L}_{\text{ts}} + \mathcal{L}_{\text{ps}}. \tag{3}$$

Training Data. As the network targets for stabilizing inpainted video frames produced by our onion-peel network, we directly uses output of the onion-peel network as input to the temporal consistency network. We use the same training images for the onion-peel network training.

2. Image Completion Result

In addition to Fig. 6 of the main paper, we provide more results for the image completion guided by reference images. In Fig. 2 - 5, we compare our method against Yu *et al*. [11] and Photoshop's content aware fill [1].

3. Video Completion Results

We provide our object removal results on the DAVIS videos [6] with shadow annotations provided by [3]. We compare our method against two state-of-the-art methods: VINet [4] and Huang *et al*. [3]. In the video file, `Video_completion.mp4`, we provide side-by-side comparisons on challenging test videos.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (TOG)*, 28(3):24, 2009. 1

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 1

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

[4] 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*, 2019. 1

[5] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *European Conference on Computer Vision*, 2018. 1

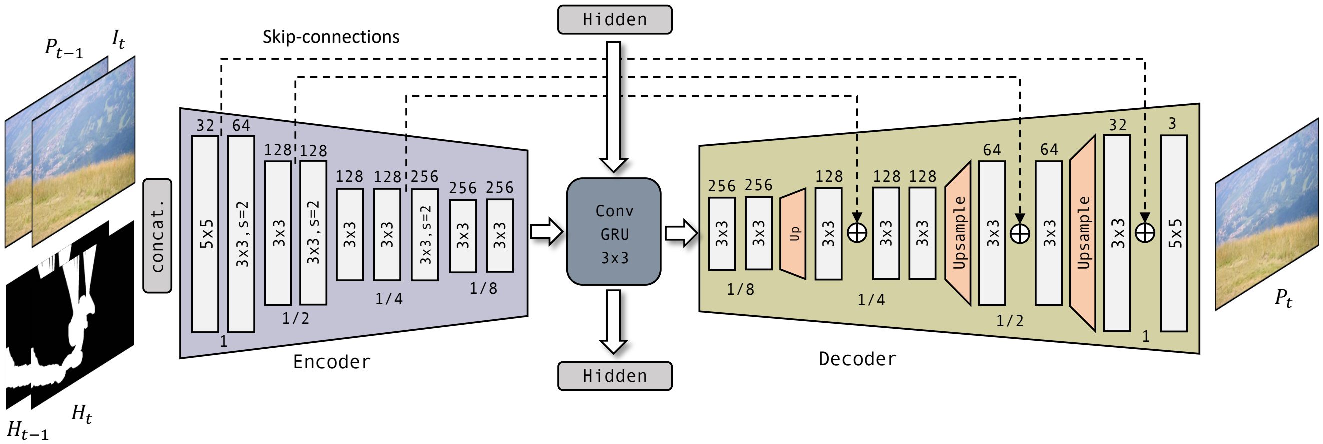

Figure 1: Temporal consistency networks. $\bigoplus$ indicates the element-wise addition.

[6] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[8] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan 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. 1

[9] Ning Xu, Linjie Yang, Dingcheng Yue, Jianchao Yang, Yuchen Fan, Yuchen Liang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. In *arXiv preprint arXiv:1809.03327*, 2018. 3, 4, 5, 6

[10] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018. 1

[11] 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*, pages 5505–5514, 2018. 1

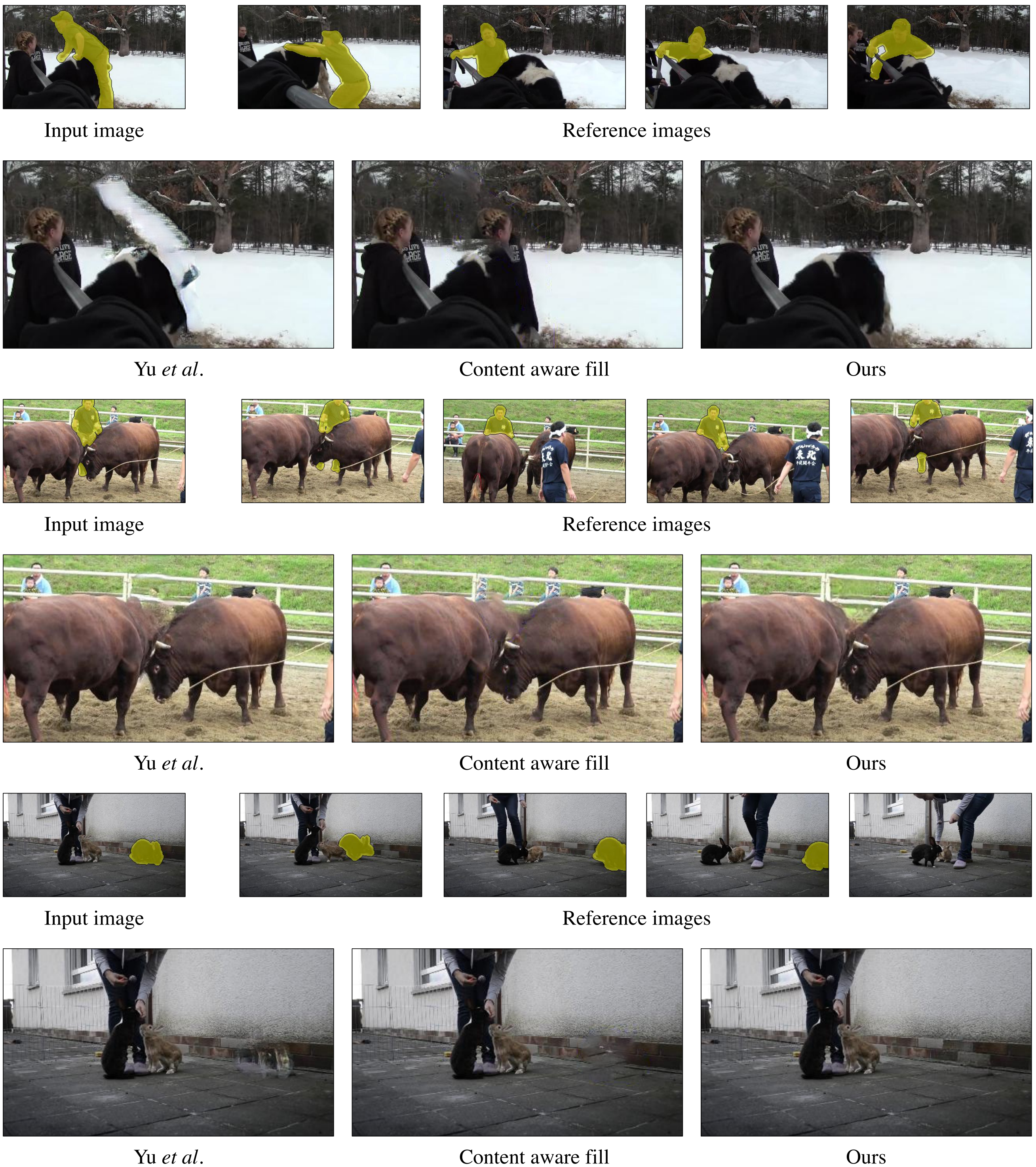

Figure 2: Examples of image completion using a group of photos (Best viewed on a high-resolution display with zoom-in). The images are from Youtube-VOS [9].

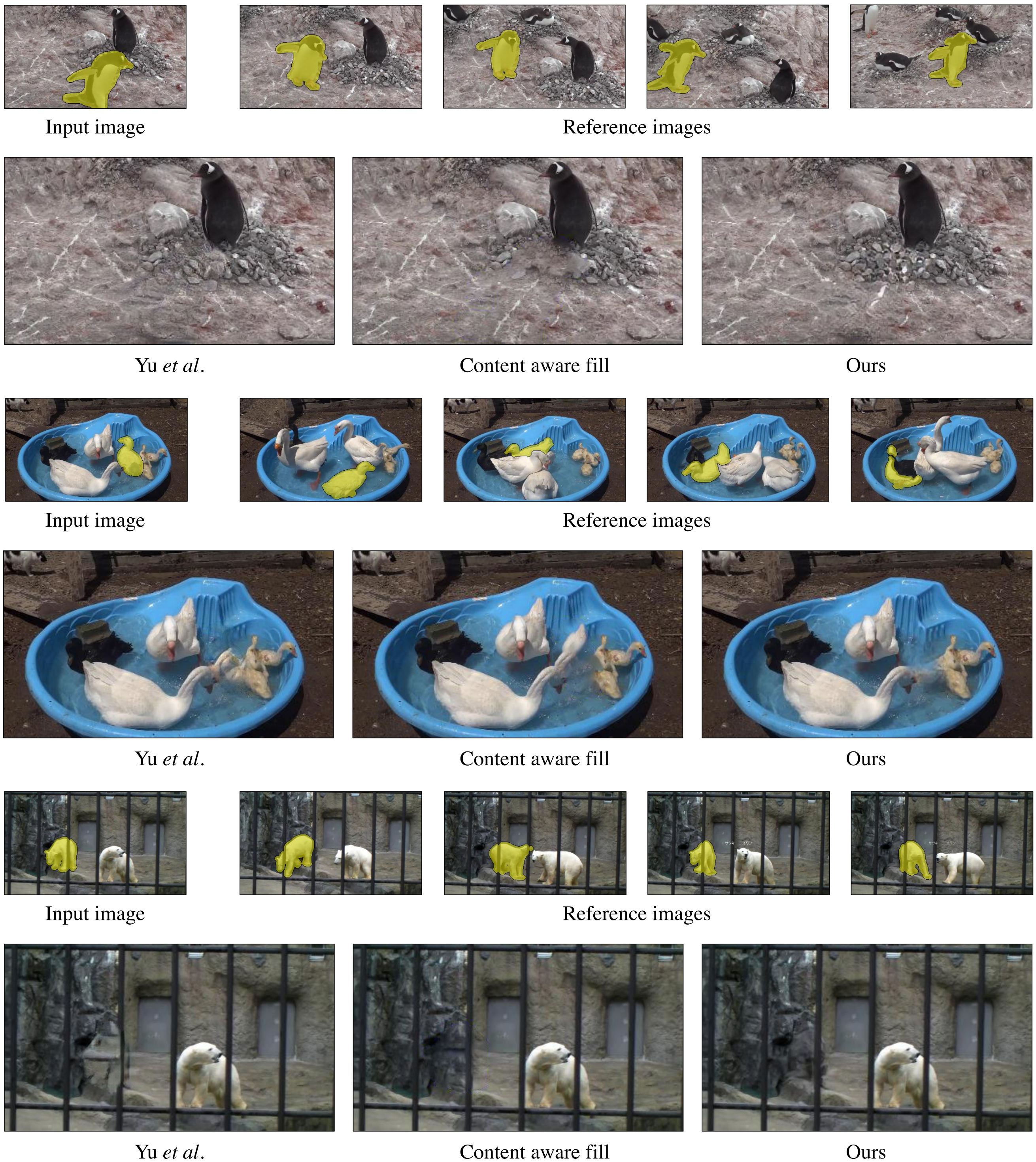

Figure 3: Examples of image completion using a group of photos (Best viewed on a high-resolution display with zoom-in). The images are from Youtube-VOS [9].

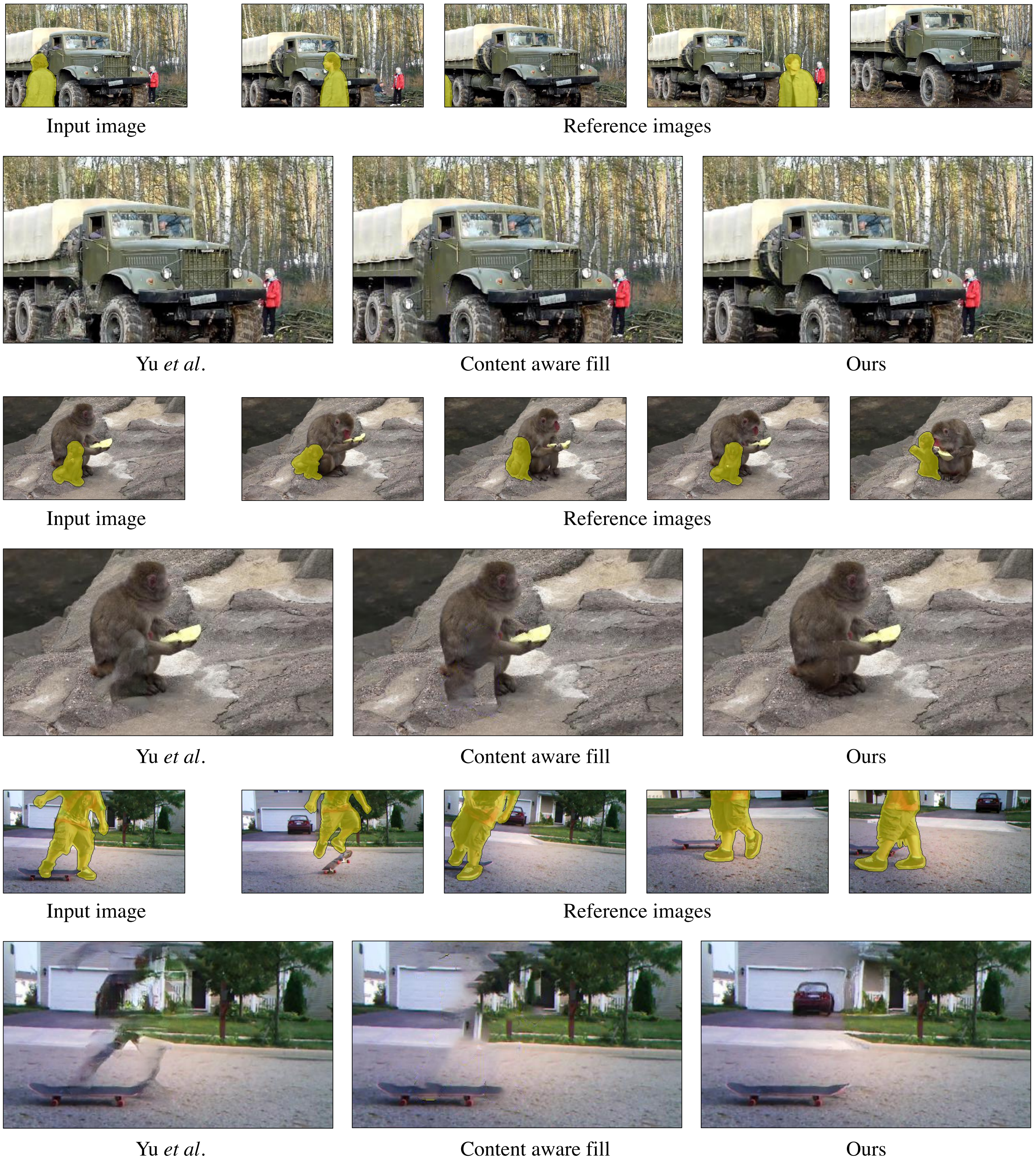

Figure 4: Examples of image completion using a group of photos (Best viewed on a high-resolution display with zoom-in). The images are from Youtube-VOS [9].

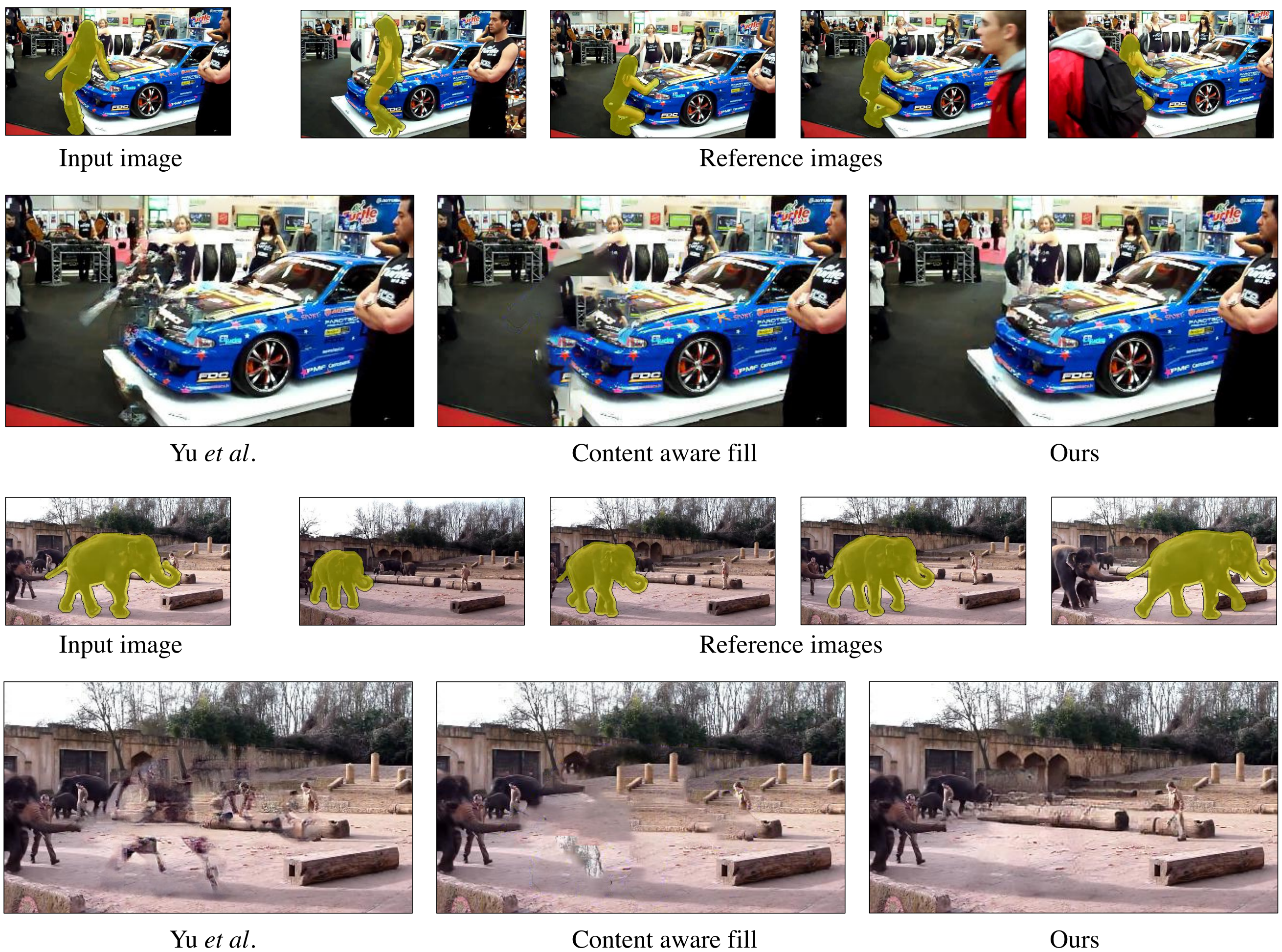

Figure 5: Examples of image completion using a group of photos (Best viewed on a high-resolution display with zoom-in). The images are from Youtube-VOS [9].