Supplementary Material for Semantic-Aware Dynamic Parameter for Video Inpainting Transformer

Eunhye Lee[1*], Jinsu Yoo[2*], Yunjeong Yang[2], Sungyong Baik[2,3], Tae Hyun Kim[1†]

{dldms1345, jinsuyoo, yunjeongyang, dsybaik, taehyunkim}@hanyang.ac.kr

[1]Dept. of Computer Science, [2]Dept. of Artificial Intelligence, [3]Dept. of Data Science, Hanyang University

A. Detailed Experimental Settings

A.1. Subcategories of super-categories

Table S1 lists the subcategories that belong to the eight super-categories we construct. COCO-Stuff dataset [1] defines a set of categories as the super-category, and we cluster them into the newly eight super-categories.

Our super-category		COCO-Stuff [1] super-category	Category
Foreground Object	**Animal**	Animal	bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe
	Vehicle	Vehicle	bicycle, car, motorcycle, airplane, bus, train, truck, boat
	Person	Person	person
	Sports	Sports	frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket
	Plant	Plant	flower, tree-merged, grass-merged
		Food	banana, apple, sandwich, orange, broccoli, hot dog, pizza, donut, cake
		Food-stuff	fruit, food-other-merged
Background	**Plain**	Ground	sand, snow, dirt-merged, gravel, road, pavement-merged, railroad, platform, playingfield
		Solid	mountain-merged, rock-merged
		Water	river, sea, water-other
		Sky	sky-other-merged
	Patterned	Building	bridge, house, roof, tent, building-other-merged
		Wall	wall-tile, wall-wood, wall-brick, wall-stone, wall-other-merged
		Window	window-blind, window-other
		Structural	fence-merged, net
		Floor	floor-wood, floor-other-merged
		Ceiling	ceiling-merged
Etc.	**Etc.**	Furniture	chair, couch, potted plant, bed, dining table, toilet
		Furniture-stuff	counter, door-stuff, light, mirror-stuff, shelf, stairs, cabinet-merged, table-merged
		Textile	rug-merged, towel, curtain, blanket, pillow, banner
		Accessory	backpack, umbrella, handbag, tie, suitcase
		Kitchen	bottle, wine glass, cup, fork, knife, spoon, bowl
		Appliance	microwave, oven, toaster, sink, refrigerator
		Electronic	tv, laptop, mouse, remote, keyboard, cell phone
		Indoor	book, clock, vase, scissors, teddy bear, hair drier, toothbrush
		Outdoor	traffic light, fire hydrant, stop sign, parking meter, bench
		Raw-material	cardboard, paper-merged

Table S1. Detailed category lists for each super-category used for our experiments.

In ablation study, the 4-expert super-categorization comprises: 'Animal/Person/Vehicle,' 'Plant/Plain Background,' 'Patterned Background,' and 'Sports/Etc.' Also, the 6-expert super-categorization includes: 'Animal,' 'Person,' 'Vehicle,' 'Plant/Plain Background,' 'Patterned Background,' and 'Sports/Etc.'

A.2. Acquiring segmentation maps for foreground object removal

We use a simple k-nearest neighbors (KNN) algorithm (`KNeighborsClassifier` function in scikit-learn [5]) for generating segmentation maps corresponding to the foreground object. The algorithm takes 30(=K) consecutive segmentation maps as inputs with holes on the masked foreground objects, and outputs a segmentation map corresponding to the mid-frame that the masked area is filled with the KNN algorithm.

A.3. Quantitative comparison on foreground object removal

Table S2 compares the Video-based Fréchet Inception Distance (VFID) [7] and temporal warping error [2] on the DAVIS [6] dataset in foreground object removal setting. Note that for the object removal task, we cannot provide the evaluation results in terms of PSNR and SSIM and the statistics of video including the unwanted objects is used to compute the VFID value due the lack of ground truth dataset for this task. The results indicate that our method generates temporally more consistent video against the baseline methods while retaining the visual quality.

Method	VFID$_{\downarrow}$	$E_{warp}(\%)_{\downarrow}$
E2FGVI [3]	0.783	0.1143
FuseFormer [4]	0.782	0.1236
SAVIT-KNN	0.781	0.0628

Table S2. Perceptual quality and temporal consistency of transformer-based video inpainting models in foreground object removal task.

B. Additional Experimental Results

B.1. More qualitative results

This section provides additional visual comparisons of SAVIT against baseline video inpainting methods on YouTube-VOS [8] and DAVIS [6] datasets for fixed region inpainting and foreground object removal tasks. In Fig. S1-S2, we compare SAVIT with existing transformer-based video inpainting networks, including STTN [9], E2FVGI [3], and FuseFormer [4], under fixed region inpainting task. The inpainted results indicate the superiority of our method in hallucinating more visually plausible images. Moreover, Fig. S3-S6 show comparisons of SAVIT against our main baseline architecture, FuseFormer [4], along with the corresponding input segmentation maps on both fixed region inpainting and foreground object removal settings. Under both settings, our SAVIT consistently synthesizes more visually pleasing content (please see results on human body and object boundaries).

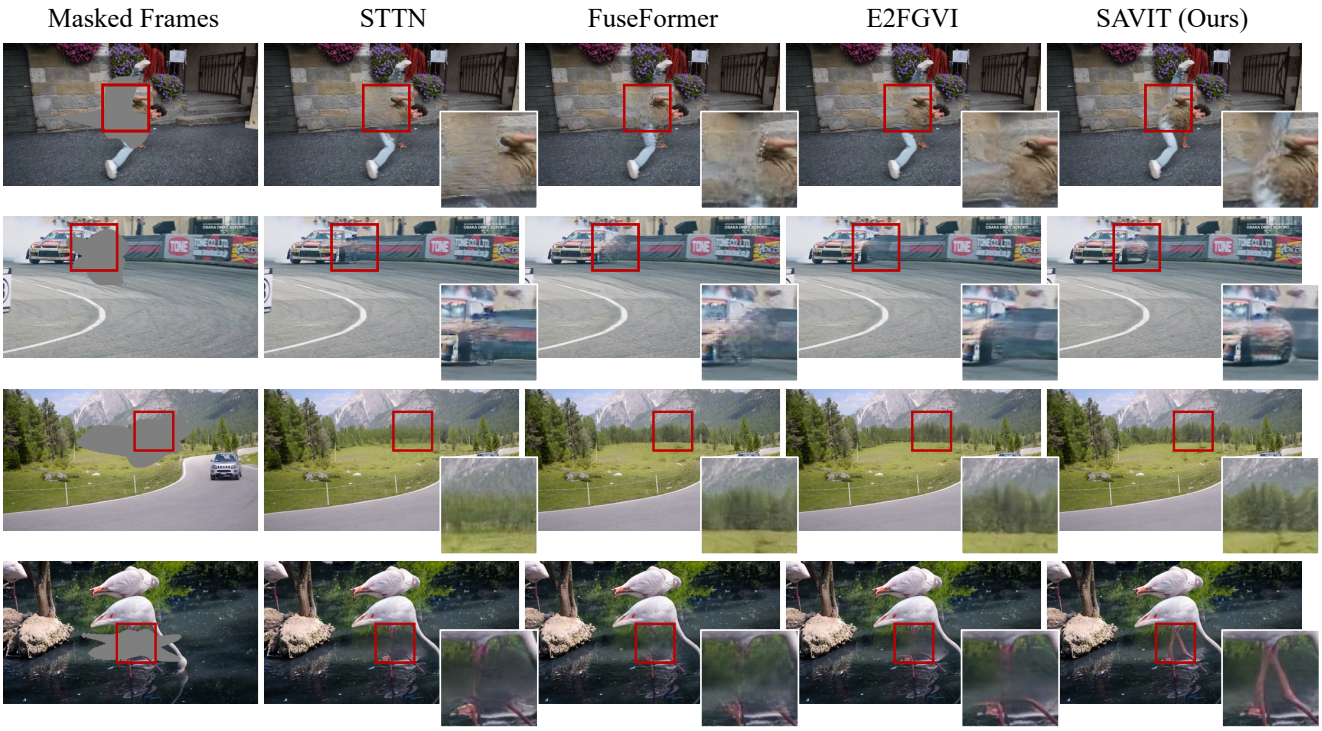

Figure S1. Qualitative comparison of SAVIT against baseline video inpainting networks on DAVIS [6] in fixed region inpainting setting.

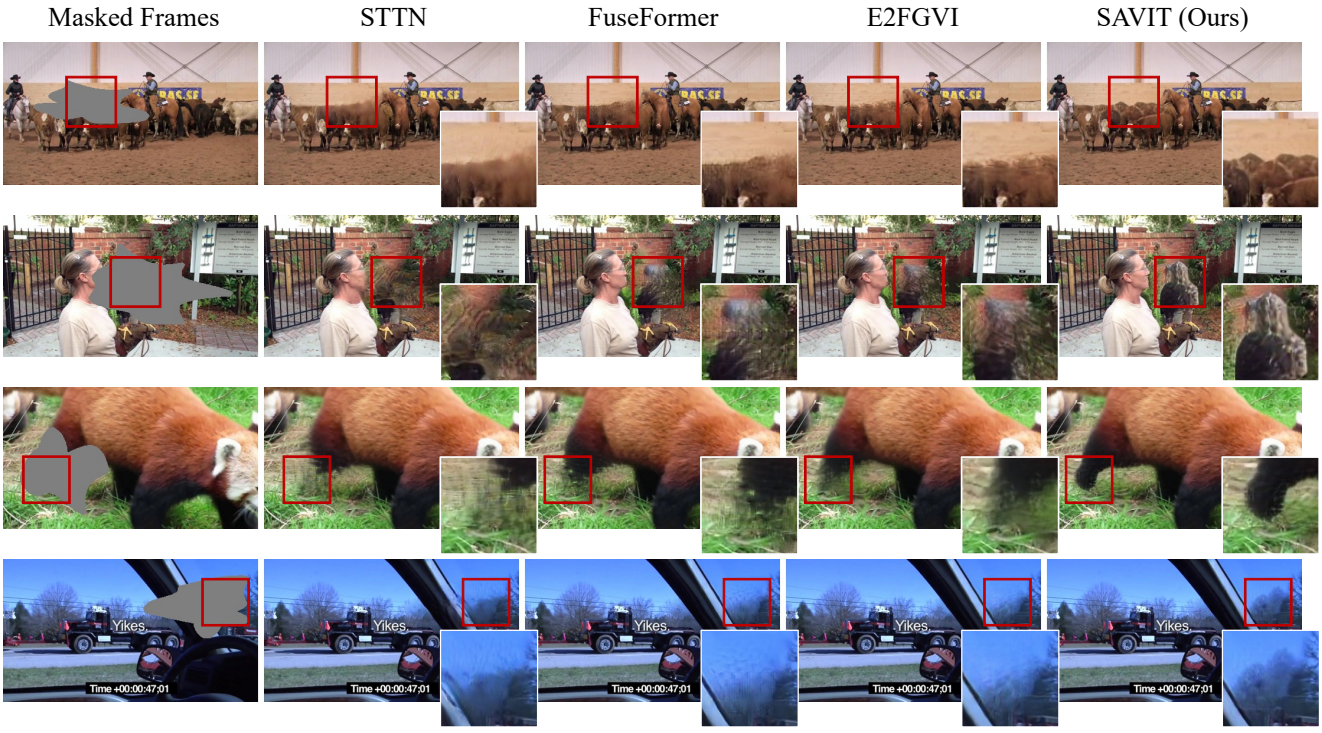

Figure S2. Qualitative comparison of SAVIT against baseline video inpainting networks on YouTube-VOS [8] in fixed region inpainting setting.

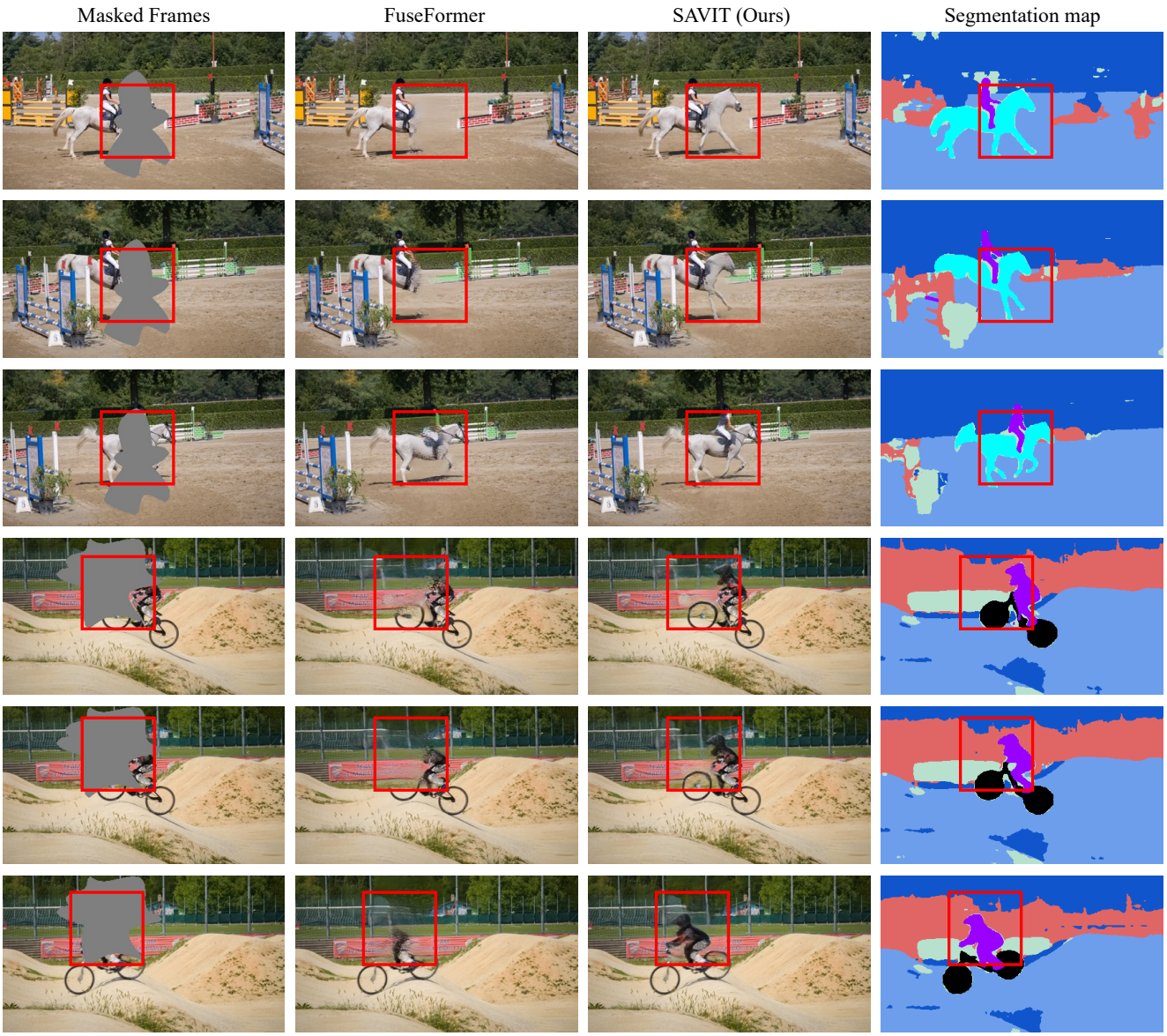

Figure S3. Input segmentation maps and qualitative results of SAVIT in comparison with FuseFormer [4] in fixed region inpainting setting.

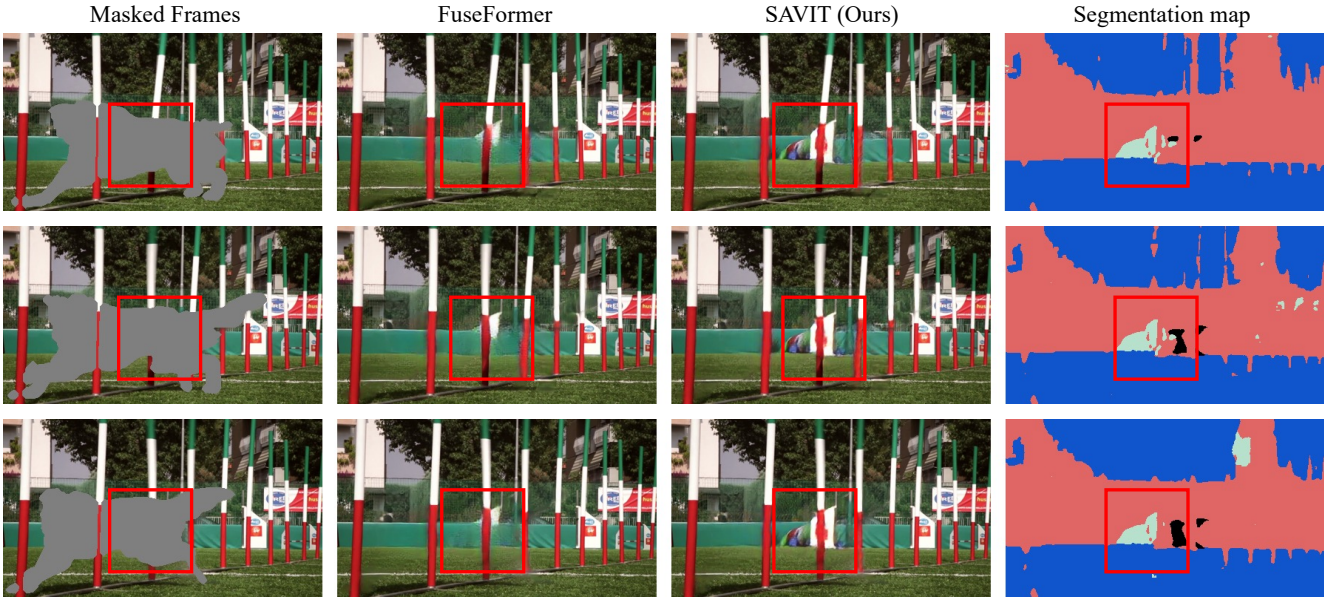

Figure S4. Input segmentation maps and qualitative results of SAVIT in comparison with FuseFormer [4] in foreground object removal setting.

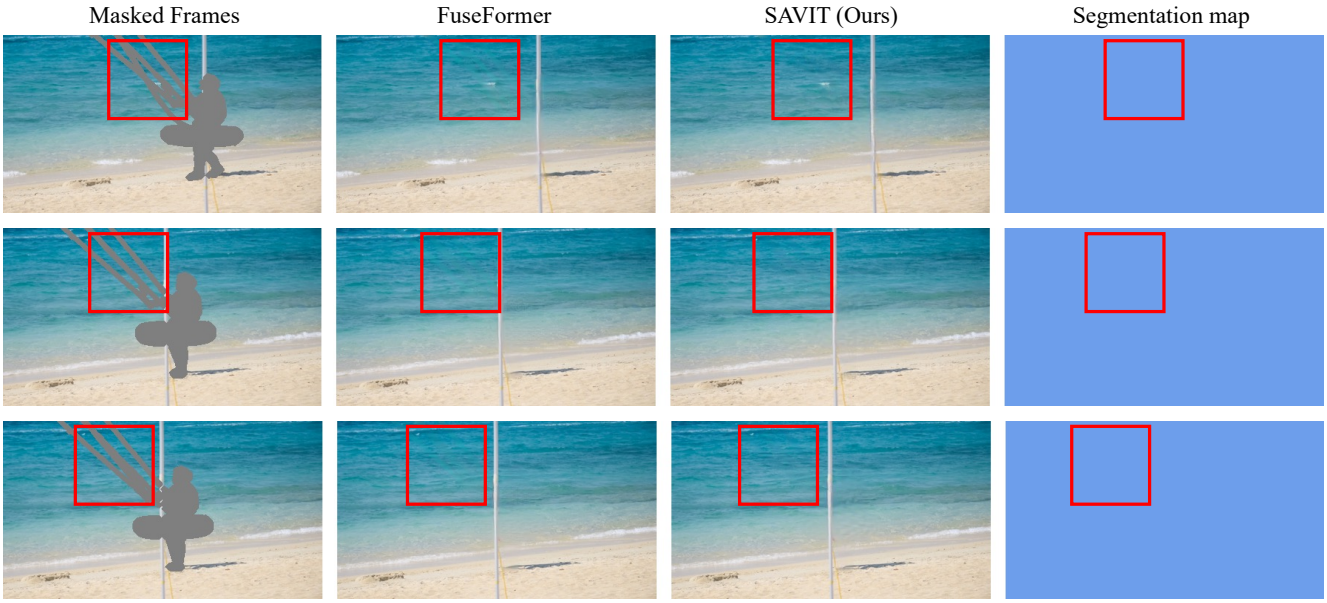

Figure S5. Input segmentation maps and qualitative results of SAVIT in comparison with FuseFormer [4] in foreground object removal setting.

Figure S6. Input segmentation maps and qualitative results of SAVIT in comparison with FuseFormer [4] in foreground object removal setting.

C. Additional ablation study

C.1. Computational cost

#experts	#param	FLOP	PSNR$^{\uparrow}$
1 (FuseFormer$_{small}$)	18.86M	290G	30.51
4 (+ Ours)	21.95M	305G	30.83
8 (+ Ours)	25.97M	320G	31.01
132 (w/o super-cat)	151.0M	778G	-

Table S3. Computational cost.

Table S3 outlines the computational cost of our method with FuseFormer$_{small}$ as a backbone. First, when comparing the first and second rows, our model effectively enhances the baseline performance by 0.3dB, while only increasing the parameters by approximately 15% and the FLOPs by 5%. Comparing the first and third rows reveals that our model significantly improves the performance (+0.5dB), with only a 10% increase in FLOPs. Notably, without our super-categorization, the model would require substantial resources (fourth row). However, we have significantly reduced the cost by using the effective super-categorization.

C.2. Dynamic block replacement

Placement	PSNR$^{\uparrow}$	SSIM$^{\uparrow}$	LPIPS$_{\downarrow}$	VFID$_{\downarrow}$
first	**31.05**	0.9615	0.0465	0.166
mid	31.01	**0.9624**	**0.0408**	**0.157**
last	30.91	0.9616	0.0437	0.157

Table S4. Dynamic block placement.

Table S4 compares various configurations that replace the original block with our dynamic block. We observe that the specific placement of our dynamic block does not yield a significant performance difference. Thus, we have chosen to place the block in the middle, and focus on demonstrating the effectiveness of leveraging semantic cues.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[2] 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. 2

[3] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[4] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 4, 5, 6

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011. 2

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[8] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2, 3

[9] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2