Breaking Temporal Consistency:
Generating Video Universal Adversarial Perturbations Using Image Models

Hee-Seon Kim, Minji Son, Minbeom Kim, Myung-Joon Kwon, Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)
{hskim98, ming0103, alsqja1754, kwon19, changick}@kaist.ac.kr

This supplementary material provides additional experimental results, such as visualized BTC-UAP and videos used in the experiments, along with further explanations that were omitted in the main paper.

S1. Supplementary Experimental settings

S1.1. Experiment

For ImageNet[1] and Kinetics 400[7], we selected images from each class using a fixed random seed(1). For UCF101[10], we created a UAP using testset1 and evaluated it on testset2 from the dataset's provided testset 1, 2, and 3.

S1.2. Discussion

As mentioned in the main paper, Figure 4 in the Discussion section presents the application of BTC-UAP generated using ResNet-101[5] and ImageNet data and evaluated against 32-frame UCF-101 videos with six different video models: SlowFast-50 (SF-50), SlowFast-101 (SF-101)[4], Temporal Pyramid Network-50 (TPN-50), Temporal Pyramid Network-50 (TPN-101) [14], NonLocal-50 (NL-50), and NonLocal-101 (NL-101) [11]. We have added more detailed settings for each baseline. ASR stands for Attack Success Rates (%).

In Figure 4-(a) of the main paper, BTC-UAP and I2V-UAP[12] were generated using the ResNet-101[5] model and ImageNet data, denoted as BTC(I) and I2V(I). BTC-UAP and I2V-UAP were also compared with those generated using ResNet-101 and UCF-101 data, denoted as BTC(V) and I2V(V). The other baselines, MI-UAP[2], DI-UAP[3], TI-UAP[13], and SI-UAP[8], were denoted as M, D, T, and S, respectively, and were generated using the ResNet-101 model and ImageNet data. Additionally, for SAP-UAP and TT-UAP generated using the video model, denoted as SAP and TT, we used the SlowFast-101 model as they exhibited the highest ASR, and UCF-101 data.

In Figure 4-(d) of the main paper, the average ASR of BTC-UAP generated using ResNet-101[5] and ImageNet data against six different models is presented. Please noted that the Temporal Similarity Loss alone cannot be used without the Adversarial Loss. Since Adversarial Loss is necessary for the attack, the optimization process began with K for the Adversarial Loss and then optimized J for the Temporal Similarity Loss.

S2. Comparison with Image-based Attack

Please note that Table S1 differs from Table 4 in the main paper solely in the dataset used for pretraining the attacking models. Due to limited space, we could not include this table in the main paper.

We used three pre-trained image models on the ImageNet dataset: ResNet101 (Res-101)[5], SqueezeNet (Squeeze)[6], and VGG16[9]. We carried out experiments to assess the transferability of Universal Adversarial Perturbations (UAPs) generated using image data and models. In Table S1, we showcase the Attack Success Rate (ASR) of adversarial videos, where the UAP is optimized on ImageNet using each corresponding image model. Among the methods compared, BTC-UAP achieved the highest average ASR, exhibiting remarkable transferability. For instance, while I2V-UAP has a total average ASR of 22.27% across all cases, BTC-UAP surpasses it with an impressive 40.82% average ASR. These findings underscore that our proposed method effectively incorporates temporal information, leading to the highest performance among image-based approaches.

S3. Videos and UAPs Visualization

In Figure S1, we visualize and display the initial 8 frames of BTC-UAP, clean, and adversarial videos. In the figure, they are denoted as BTC-UAP, Clean, and Adv, respectively. Although BTC-UAP is originally constrained by a value of 16/255, it has been normalized between 0 and 1 for better visualization. The clean video refers to the UCF101[10] original video, while the adversarial video represents the video with BTC-UAP added. It can be observed that the overall appearance of the video is not significantly disrupted.

Source Models	Attack	Target Models SF-50 (UCF-101)	SF-101 (UCF-101)	TPN-50 (UCF-101)	TPN-101 (UCF-101)	NL-50 (UCF-101)	NL-101 (UCF-101)	AVG.
Res-101 (ImageNet)	MI-UAP	15.35	9.29	11.73	7.95	19.31	18.40	13.67
	DI-UAP	15.75	10.04	11.52	9.16	18.67	18.59	13.95
	TI-UAP	21.83	14.92	16.90	14.27	31.20	29.97	21.51
	SI-UAP	21.69	17.57	22.52	20.35	35.30	34.52	25.33
	I2V-UAP	25.60	18.45	42.55	29.16	25.82	41.27	30.48
	BTC-UAP	**49.01**	**36.98**	**65.37**	**47.67**	**49.41**	**63.34**	**51.96**
VGG16 (ImageNet)	MI-UAP	14.17	8.89	9.99	6.91	22.34	22.36	14.11
	DI-UAP	17.25	10.50	12.05	8.54	21.99	21.75	15.35
	TI-UAP	22.79	15.40	15.69	11.81	31.36	30.45	21.25
	SI-UAP	17.76	12.29	13.28	9.56	26.41	25.55	17.47
	I2V-UAP	17.84	11.03	15.43	10.44	16.47	21.53	15.46
	BTC-UAP	**37.90**	**31.07**	**43.30**	**34.47**	**51.45**	**58.89**	**42.85**
Squeeze (ImageNet)	MI-UAP	14.65	9.45	10.07	6.96	21.48	17.97	13.43
	DI-UAP	13.98	8.30	9.91	6.67	20.46	19.68	13.17
	TI-UAP	**26.73**	19.84	19.74	16.07	**43.28**	**39.26**	27.49
	SI-UAP	15.16	9.80	10.85	7.95	20.70	17.49	13.66
	I2V-UAP	19.04	17.01	20.14	14.09	27.08	27.93	20.88
	BTC-UAP	24.83	**20.57**	**24.21**	**19.26**	38.97	38.00	**27.64**

Table S1: **Attack success rates (%) of UAPs generated on image models using image data.** UAPs are optimized on ImageNet and adversarial videos are generated by adding UAPs to UCF101 videos. The bold numbers indicate the highest attack success rate among attack methods.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[2] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[3] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[4] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. 2017.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[8] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[11] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[12] Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Cross-modal transferable adversarial attacks from images to videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15064–15073, 2022.

[13] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.

[14] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020.

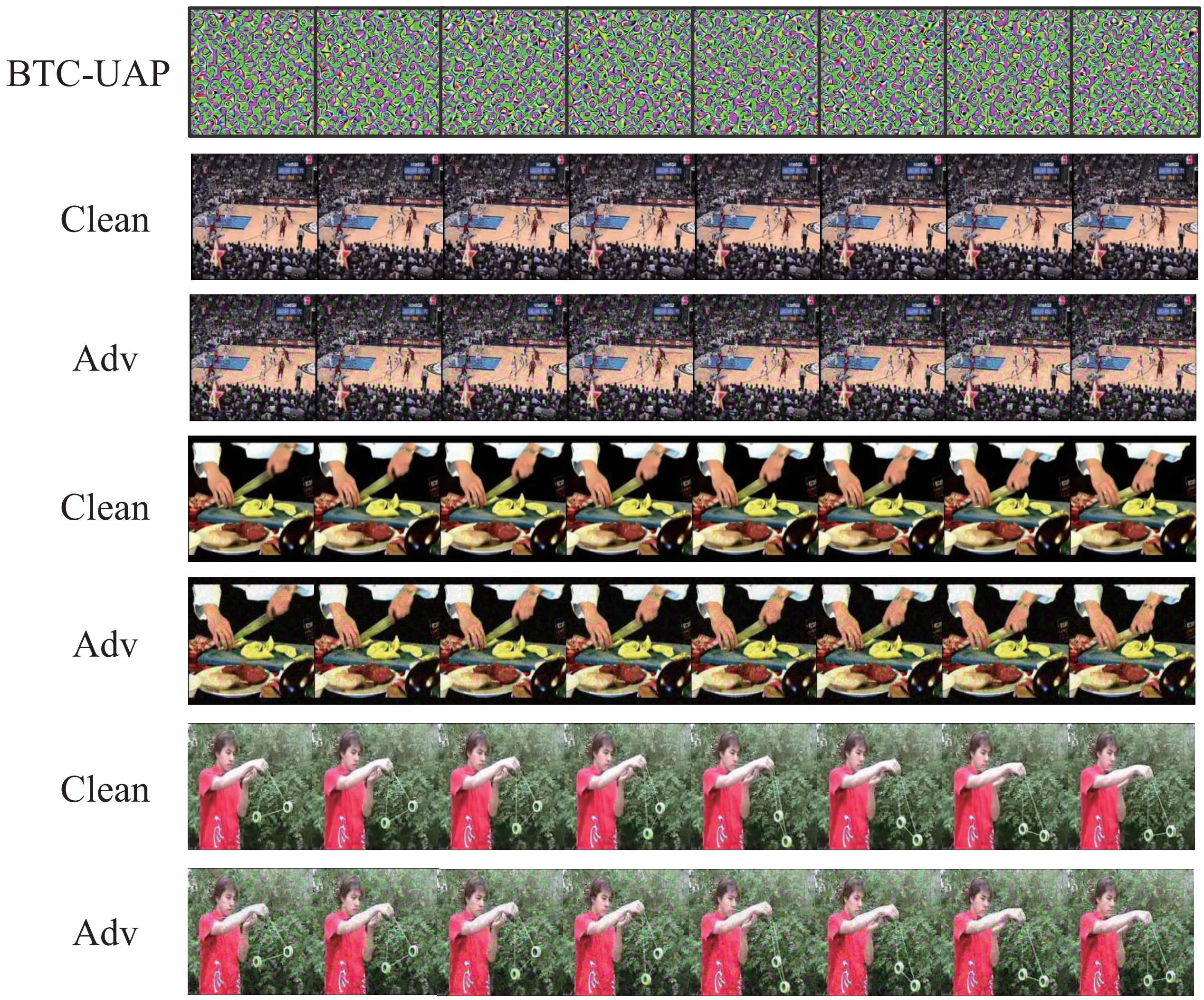

Figure S1: **Videos and UAPs visualization.** We visualize and display the initial 8 frames of BTC-UAP, clean, and adversarial(Adv) videos.