Supplementary Material for "CLIP-Guided Generative Networks for Transferable Targeted Adversarial Attacks"

Hao Fang[1†], Jiawei Kong[2†], Bin Chen[2#], Tao Dai[3], Hao Wu[4], and Shu-Tao Xia[1]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Harbin Institute of Technology, Shenzhen [3] Shenzhen University
[4] Shenzhen Digital Certificate Authority CO., Ltd
fang-h23@mails.tsinghua.edu.cn, kongjiawei@stu.hit.edu.cn
chenbin2021@hit.edu.cn, daitao.edu@gmail.com, whpc79@163.com,
xiast@sz.tsinghua.edu.cn

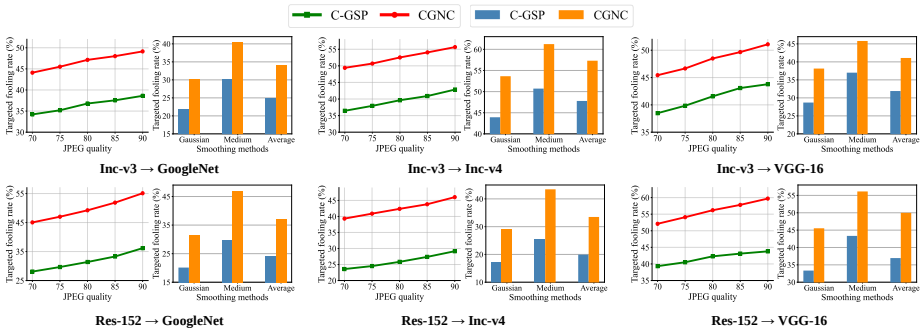

Fig. 1: Targeted transferability comparison of our CGNC and C-GSP [10] on different victim models under various input processing defenses.

A Additional Experiments

We present additional experimental results to conduct a comprehensive comparison and in-depth analysis. Similarly, we follow [2, 10] and adopt their used 8 classes as our target categories and compute the average attack success rates (ASR) on the 8 target classes as metrics. Unless stated otherwise, we use the ImageNet-NeurIPS (1k) dataset [6] to evaluate the attack performance.

[†]Equal contribution.
[#]Corresponding author.

Table 1: Attack success rates (%) for multi-target attacks against regularly trained models on ImageNet validation set. * represents white-box attacks.

Source	Method	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogleNet	VGG-16
Inc-v3	C-GSP	84.25*	45.34	35.99	36.70	57.29	41.88	48.54
	CGNC	**96.59***	**57.82**	**46.84**	**44.13**	**65.90**	**53.40**	**56.27**
Res-152	C-GSP	34.92	33.18	18.43	88.65*	62.61	41.41	44.55
	CGNC	**56.00**	**50.37**	**32.26**	**96.44***	**86.69**	**63.84**	**63.90**

Table 2: Comparison results on three black-box models under different perturbation budgets ϵ. The surrogate model is Res-152.

Method	VGG-16			Inc-v3			DN-121		
	8/255	12/255	16/255	8/255	12/255	16/255	8/255	12/255	16/255
Logit	2.71	5.91	9.20	1.65	4.70	10.10	2.86	6.62	12.70
SU	3.55	9.13	14.28	2.34	6.59	12.36	3.95	9.62	16.13
C-GSP	15.48	32.11	45.90	10.43	23.98	37.70	31.66	56.79	64.20
Ours	**21.46**	**46.28**	**63.36**	**15.04**	**37.35**	**53.39**	**45.83**	**73.05**	**85.66**

A.1 Evaluation under Input Processing Defenses

As mentioned before, we provide results on more victim models to compare our method and C-GSP [10] under various input processing defenses. Figure 1 verifies that our CGNC consistently surpasses C-GSP under the considered defense strategies, revealing the effectiveness of our CLIP-empowered network.

A.2 Evaluation on ImageNet Validation Set

For a more overall analysis, we compare our proposed CGNC and C-GSP [10] on the whole ImageNet [1] validation set (50k samples). The experimental results are shown in Table 1. Evidently, our method stably achieves better transferability, with average improvements of 19.66% and 9.77% in black-box ASR using Res-152 and Inc-v3 as surrogate models respectively.

A.3 Evaluation on Different Perturbation Budget ϵ

We then explore attacks under different ϵ values. Specifically, we additionally consider smaller ϵ values of $8/255$ and $12/255$, where the adversarial perturbations are more imperceptible. The experimental results in Table 2 reveal that our proposed network outperforms both the powerful iterative attacks Logit [11], SU [9], and the state-of-the-art (SOTA) multi-target generative attack C-GSP.

A.4 Ablation analysis of CGNC

In this section, we use Res-152 as the substitute model and present additional ablative experiments concerning our proposed CGNC to verify the contribution of each technique and investigate the influence of certain hyper-parameters.

Table 3: ASR of CGNC and its three variants. * denotes white-box attacks.

Architecture	VGG-16	GoogleNet	Inc-v3	Res-152	DN-201
CGNC	63.36	62.23	53.39	95.85*	82.69
CGNC-P	49.84	47.76	44.15	91.18*	71.09
CGNC-F	56.85	54.80	52.14	96.45*	82.19
CGNC-t	50.55	50.49	44.55	91.30*	73.38

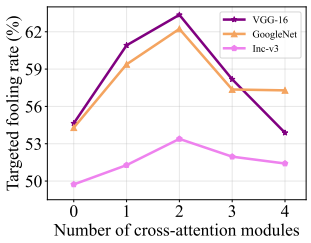

Fig. 2: ASR on three target models with various numbers of cross-attention modules.

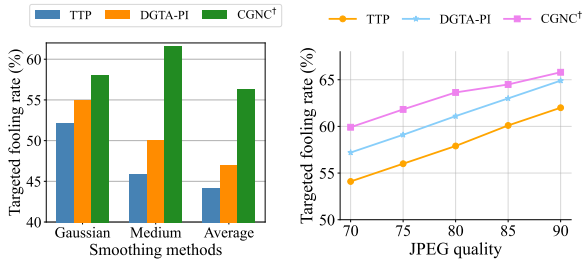

Fig. 3: Comparison of our single-target variant CGNC† with the SOTA single-target attacks under various input processing defenses. The victim model is VGG-16.

The effect of VL-Purifier. We first explore the influence of the VL-Purifier module. Specifically, we design CGNC-P that removes the VL-Purifier from the CGNC network. From Table 3, we find that directly incorporating CLIP's text embedding into the generator leads to serious performance degradation, which confirms the importance of this purifier module.

The effect of feature fusion. To verify the effectiveness of feature fusion operation in the F-Encoder, we introduce a variant CGNC-F which cancels the concatenate operation for feature fusion. The experimental results in Table 3 validate the significance of the multi-modal feature fusion process.

The effect of CLIP's text embedding. We analyze the effect of the text embedding by implementing a version CGNC-t that replaces all the text inputs with one-hot labels. The remarkable improvement from CGNC-t to CGNC shown in Table 3 directly confirms the effectiveness of incorporating text information into the generator's architecture.

Numbers of the cross-attention modules. We analyze the impact of the number of cross-attention modules on the attack performance. As illustrated in Fig. 2, the generator exhibits optimal performance across all considered target models when employing two cross-attention modules. Consequently, we integrate two cross-attention modules into the backbone of the CA-Decoder.

Scales of training data. We adopt the same settings as previous generative attacks (*e.g.*, CD-AP [4], TTP [5], C-GSP [10], and DGTA-PI [2]) and thus use the whole ImageNet training set to train generators. To investigate the influence of amount of training data, we further conduct training with different numbers

Table 4: ASR under various proportions of ImageNet training set.

Datset proportion	1/4	1/2	3/4	1
C-GSP (Res-152)	25.65	28.99	38.21	40.52
CGNC (Res-152)	**37.88**	**45.79**	**51.20**	**58.40**

of images. Tab. 4 shows that the scale of the training set indeed has a notable influence on the performance and our CGNC always outperforms C-GSP [10].

A.5 More Comparison with Single-Target Attacks

We provide more experimental results regarding Res-152 as the surrogate model to compare the single-target variant CGNC† obtained through masked fine-tuning (MFT) with SOTA single-target methods, including GAP [7], CD-AP [4], TTP [5], and DGTA-PI [2].

Comparison under Defense Strategies. We consider the same defense strategies discussed in the main body of this manuscript. On attacking the adversarially robust model, our method achieves a notable average improvement of 4.37% across six target models as shown in Table 5, demonstrating the excellent generalization ability of the proposed CGNC†.

For input defense strategies, Fig. 3 shows that our CGNC† also outperforms other methods when targeting models equipped with such defenses, especially for the input smoothing operations. It is also noteworthy that our method, which initially lags behind DGTA-PI [2] when attacking normally trained VGG-16, achieves a comprehensive lead after applying the smoothing operations and JPEG compression, highlighting the robustness and superiority of CGNC† in handling various input-based defenses.

These results again indicate that although CGNC is designed for multi-target attacks, it can achieve better performance than these powerful single-target attack methods by simply fine-tuning it with a mask operation, revealing its great potential and scalability.

Table 5: Comparison of the proposed CGNC† with existing single-target attacks against target models with robust training mechanisms.

Method	Inc-v3$_{adv}$	IR-v2$_{ens}$	Res50$_{SIN}$	Res50$_{IN}$	Res50$_{fine}$	Res50$_{Aug}$
GAP	5.72	4.51	7.33	71.04	83.64	52.07
CD-AP	3.77	6.48	7.09	63.72	76.79	49.67
TTP	27.99	26.08	24.61	72.47	74.51	70.96
DGTA-PI	31.10	30.07	27.70	77.13	80.55	**76.78**
CGNC†	**31.55**	**33.63**	**33.31**	**88.34**	**89.74**	72.96

Ablation analysis of the masked fine-tuning. To further verify the effectiveness of the proposed mask fine-tuning mechanism, we conduct ablation

Table 6: ASR of 8 different target classes. We compare the normal fine-tuning and our masked fine-tuning technique (*i.e.*, CGNC†) for single-target attacks.

Source	Method	Target class id							
		150	426	843	715	952	507	590	62
Res-152	CGNC	72.10	46.02	60.08	50.97	60.63	54.78	47.03	75.58
	Fine-tuning	73.63	56.43	71.57	45.78	70.82	59.25	45.97	75.43
	MFT	**78.38**	**63.32**	**76.12**	**56.47**	**78.40**	**64.18**	**49.65**	**84.20**
Inc-v3	CGNC	64.27	46.17	47.78	38.82	60.32	52.65	**51.05**	61.63
	Fine-tuning	70.23	61.72	72.43	48.30	64.43	68.12	42.65	56.03
	MFT	**81.63**	**72.20**	**81.82**	**52.38**	**77.52**	**73.07**	49.13	**72.22**

Table 7: ASR on ViT-based models. The surrogate is Res-152.

Method	ViT-B/16	CaiT-S/24	Visformer-S	DeiT-B	LeViT-256	TNT-S
C-GSP [10]	11.78	32.00	36.60	35.58	37.85	31.00
CGNC	**19.46**	**54.56**	**58.70**	**59.90**	**57.53**	**48.40**

experiments and calculate the average ASR for each target class across the six black-box models. The results in Table 6 illustrate the significance of both fine-tuning and patch-wise mask operation.

A.6 Attacks on Transformer-based models.

We also evaluate on six ViT-based models in Tab. 7. The results reveal that our CGNC also consistently exhibits better performance than C-GSP [10] on Transformer-based models.

B Limitations & Future work.

In this paper, we adopt a simple yet effective text template "a photo of a {class}" recommended by CLIP [8], which has been proven effective in a variety of tasks. However, due to the excessive reliance [3] on the statistical features of 'photo', this text template may limit the transferability performance to a certain extent, particularly for target datasets with stylized images. Future research can consider introducing more accurate or detailed text as the description of the target class, such as the recommended list of eighty templates of text prompt by CLIP [8], *e.g.*, "a sculpture of a {}", "an art of {}". They can use some of their averaged representations as the generic representations of the target class. Another promising approach is to choose a related pre-training task (*e.g.*, classification) and use prompt learning [12] to acquire the representation of the target category. These learned prompts can better represent the target class and distinguish features from different categories.

C More Visualization

We provide more visualization results of generated perturbations and adversarial samples in Fig. 4. The generated perturbations carry rich semantic patterns of the target class, and as we change the input text condition, the generated patterns vary accordingly to the target class. This once again demonstrates the effectiveness of our method in modeling the target features, as well as the success of conditioning the generator with CLIP's text embeddings.

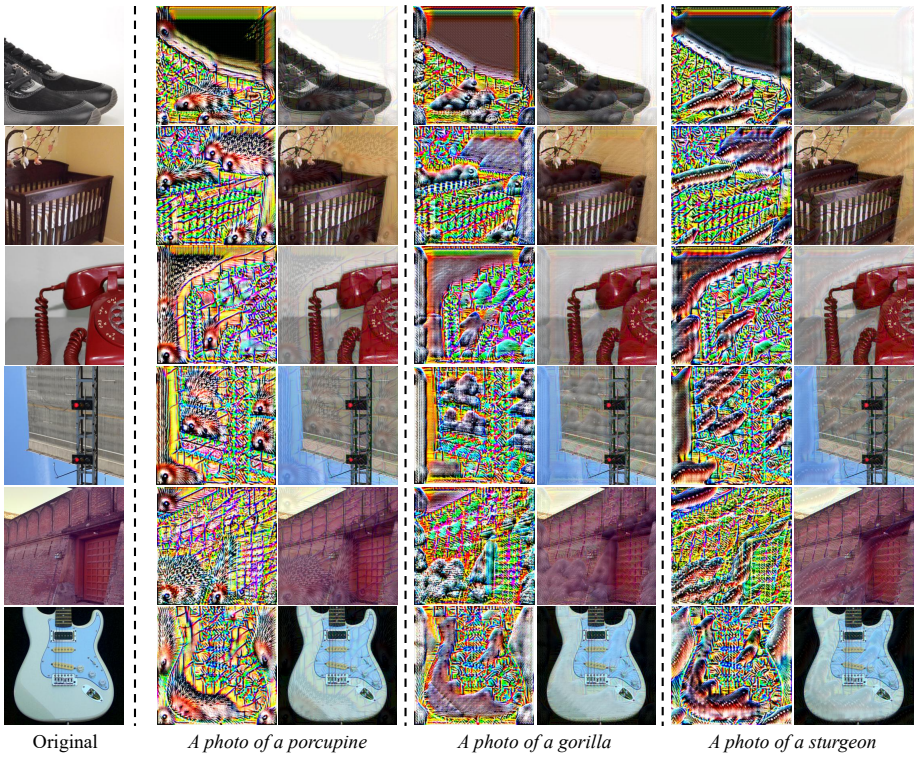

Fig. 4: Visualization of the generated perturbations and adversarial samples.

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)

2. Feng, W., Xu, N., Zhang, T., Zhang, Y.: Dynamic generative targeted attacks with pattern injection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16404–16414 (2023)
3. Huang, Z., Zhou, A., Ling, Z., Cai, M., Wang, H., Lee, Y.J.: A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11685–11695 (2023)
4. Naseer, M.M., Khan, S.H., Khan, M.H., Shahbaz Khan, F., Porikli, F.: Cross-domain transferability of adversarial perturbations. Advances in Neural Information Processing Systems **32** (2019)
5. Naseer, M., Khan, S., Hayat, M., Khan, F.S., Porikli, F.: On generating transferable targeted perturbations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7708–7717 (2021)
6. NeurIPS: `https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack/data`. Kaggle, 2017
7. Poursaeed, O., Katsman, I., Gao, B., Belongie, S.: Generative adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4422–4431 (2018)
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
9. Wei, Z., Chen, J., Wu, Z., Jiang, Y.G.: Enhancing the self-universality for transferable targeted attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12281–12290 (2023)
10. Yang, X., Dong, Y., Pang, T., Su, H., Zhu, J.: Boosting transferability of targeted adversarial examples via hierarchical generative networks. In: European Conference on Computer Vision. pp. 725–742. Springer (2022)
11. Zhao, Z., Liu, Z., Larson, M.: On success and simplicity: A second look at transferable targeted attacks. Advances in Neural Information Processing Systems **34**, 6115–6128 (2021)
12. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)