CLIP-Guided Generative Networks for Transferable Targeted Adversarial Attacks

Hao Fang[1†], Jiawei Kong[2†], Bin Chen[2,3,4#], Tao Dai[5], Hao Wu[6], and Shu-Tao Xia[1,3]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Harbin Institute of Technology, Shenzhen [3] Pengcheng Laboratory
[4] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[5] Shenzhen University [6] 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

Abstract. Transferable targeted adversarial attacks aim to mislead models into outputting adversary-specified predictions in black-box scenarios. Recent studies have introduced *single-target* attacks that train a generator for each target class to generate highly transferable perturbations, resulting in substantial computational overhead when handling multiple classes. *Multi-target* attacks address this by training only one class-conditional generator for multiple classes. However, the generator simply uses class labels as conditions, failing to leverage the rich semantic information of the target class. To this end, we design a **C**LIP-guided **G**enerative **N**etwork with **C**ross-attention modules (CGNC) to enhance multi-target attacks by incorporating textual knowledge of CLIP into the generator. Extensive experiments demonstrate that CGNC yields significant improvements over previous multi-target attacks, e.g., a 21.46% improvement in success rate from Res-152 to DenseNet-121. Moreover, we propose the masked fine-tuning to further strengthen our method in attacking a single class, which surpasses existing single-target methods.

1 Introduction

With the rapid progress of deep learning, deep neural networks (DNNs) have been widely applied in many security-critical fields, such as autonomous driving [15, 37], financial systems [53], and point cloud modeling [69, 70]. However, DNNs are corroborated to be vulnerable to adversarial attacks [20, 26, 58], which attempt to fool models with adversarial examples crafted by adding imperceptible perturbations to the original inputs. Based on the attack goal, adversarial attacks can be categorized into untargeted and targeted attacks. Untargeted attacks attempt to fool DNNs into predicting incorrect labels while targeted attacks aim at triggering the model to output the attacker-desired predictions.

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

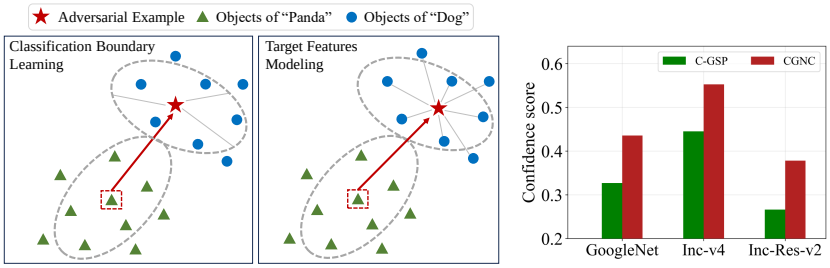

Fig. 1: (a) Targeted attacks from 'Panda' to 'Dog'. Fig. (a) illustrates that previous multi-target methods [27,67] generate perturbations simply conditioned by class indices or one-hot vectors and only learn the classification boundary specific to the surrogate model. In contrast, CGNC exploits CLIP's meaningful guidance to effectively capture the feature distribution inherent to the target data. (b) We directly feed *the scaled perturbations* generated by our CGNC and C-GSP [67] into three black-box classifiers and reveal that our perturbations achieve significantly higher mean confidence of the target class, demonstrating its superiority in modeling the target feature distribution.

Recent investigations into the adversarial transferability [2, 3, 44, 58] have demonstrated that adversarial examples crafted for a white-box surrogate model can also mislead other unseen black-box models. Since this attack does not require access to the target model, it exposes a serious security threat to real-world applications of DNNs and motivates a wide range of studies [1, 5, 42, 46]. Despite the remarkable performance on *untargeted* transferable attacks, these approaches produce unsatisfactory results for *targeted* attacks due to their over-reliance on the white-box surrogate. Existing studies on transferable targeted attacks can be categorized into *instance-specific* [11, 15, 23, 39, 43, 60, 66] and *instance-agnostic* attacks [19, 37, 44, 48, 49, 64]. Specifically, instance-specific attacks [11, 12, 65] iteratively perform gradient updating to craft adversarial perturbation tailored to a specific natural sample, yet producing low targeted transferability due to overfitting [67]. Conversely, instance-agnostic attacks learn a universal perturbation [45, 71] or a perturbation generator [48, 51] based on data distribution rather than the specific instance, alleviating the data-specific over-fitting issues and achieving a higher adversarial transferability. Recent studies further explore generative attacks to produce highly transferable perturbations, which can be divided into *single-target* and *multi-target* attacks as follows.

Single-target attacks [19, 48, 49, 62] exhibit impressive performance by training perturbation generators for target categories. However, they require training a generator for each target, which can lead to a heavy computation burden when attacking numerous classes, and thus are not applicable in real-world classification systems that usually contain hundreds/thousands of target classes [27]. To address this, [27, 67] propose multi-target attacks that train a single conditional generator for multi-target classes. By specifying the desired label as conditioning input, the trained generator can efficiently generate the corresponding adversar-

ial perturbation. Nevertheless, these methods simply adopt class indices [27] or one-hot vectors [67] of the target labels as conditions, and thus solely rely on the classification information from the surrogate model as the guidance of the target category, hence resulting in only modest black-box fooling rates.

In this paper, we build upon the research line of multi-target attacks and propose a novel CLIP-guided Generative Network with Cross-attention modules (CGNC) to boost the attack effect. Concretely, we revisit the architecture of the conditional generators used in [27,67] and argue that the simple class conditions, *e.g.*, class indices or handcrafted one-hot vectors, limit the transferability of the generated perturbations (see Fig. 1(a) for comparison). Motivated by the impressive effects of CLIP's encoded text embeddings in multi-modal learning, we introduce concise text descriptions of the target classes to encode them as class-specific representations encapsulated with abundant information, which assist the generator in learning target class distribution and ultimately lead to a fundamental transferability improvement. To better incorporate the text information, we improve the condition-inputting mechanism by adding the cross-attention layers that have been proven effective in models with various input modalities. Results in Fig. 1(b) confirm the capability of our CGNC in capturing the target distribution. Besides, we propose a masked fine-tuning (MFT) technique, which fine-tunes the trained conditional generator with a fixed text condition of the desired class for further improvement in attacking a single class.

With all the above efforts, our method achieves great efficiency and scalability. When there are hundreds of target classes and single-target methods become impractical, the proposed conditional network can be utilized to achieve significant performance over previous multi-target attacks. Conversely, for scenarios involving only a few target classes, the proposed MFT mechanism enhances CGNC to outperform existing single-target methods while substantially reducing computational costs. In summary, our main contributions are as follows:

- We propose CGNC, a novel CLIP-guided generative network with cross-attention layers that fully exploits the textual knowledge provided by the advanced CLIP model for enhanced multi-target attacks.
- We introduce a masked fine-tuning mechanism to improve the single-target performance by adapting the CGNC to an individual target class.
- Extensive experiments show that CGNC remarkably enhances the targeted transferability compared to previous attack methods in a range of settings.

2 Related Work

2.1 Vision-Language Models

Vision-language models (VLM) drawn great attention [21,22] due to their promising potential in learning general visual and textual representations through contrastive pre-training on large-scale image-text pairs. Given the powerful capacity of text descriptions in modeling multi-modal tasks, these models can be effectively adapted to diverse downstream tasks through appropriately formulated

textual prompts, including video understanding [36], image manipulation [7], and text-to-image synthesis [59]. Inspired by these works, we propose to harness CLIP-based multi-modal learning to empower our perturbation generator, facilitating more effective learning of the target feature distribution.

2.2 Adversarial Attacks

Among various security threats to DNNs [8, 16, 18, 68], the adversarial attack is one of the most formidable and well-known one, which can be classified into *instance-specific attacks* and *instance-agnostic attacks* [67].

Instance-specific Attacks. Since the pioneering work [58] highlighted the vulnerability of neural networks, numerous gradient-based optimization methods [3, 13, 26, 38, 66] have been proposed to craft image-dependent perturbations. MIM [11] integrates a momentum item into the gradient update for utilizing the previous gradient information to avoid plunging into poor local optimum. DIM [65] enhances the transferability by randomly diversifying the input pattern, and TIM [12] implements the attack by convolving the gradient with a predefined kernel. In addition, intermediate feature space [35, 63] and classifier information [34] are exploited to enhance attack effects, while [72] leverages logit-based loss to achieve competitive results. More advanced works [6, 28, 30] consider an ensemble of multiple surrogate models to reduce over-fitting.

Instance-agnostic Attacks. In contrast to instance-specific attacks, instance-agnostic attacks learn a universal perturbation [17, 45] or a generative model [44, 48, 49, 51, 61, 67] for crafting adversarial examples. By modeling the global data distribution, these methods have shown better transferability and attracted more attention in recent years. Specifically, UAP [45] and AAA [47] learn a universal perturbation to fool the model based on concrete data and compressed impression, respectively. Many subsequent works, such as GAP [51], focus on boosting the transferability of adversarial attacks by training generative models. These generative attacks can be categorized into the following two types.

Single-target Attacks. This type of attack requires training a generator for each target class. [64] first introduce the generative adversarial networks (GAN) [25] to generate adversarial perturbations. Then, CD-AP [48] concentrates on the domain-invariant adversaries and launches highly transferable cross-domain attacks using a relativistic supervisory. TTP [49] modifies the loss function and proposes to match the target distribution to mitigate over-fitting to the surrogate model. Subsequent methods achieve better transferability based on a dynamic network with pattern injection [19] or a feature discriminator [62].

Multi-target Attacks. MAN [27] notes that when dealing with numerous classes, single-target attacks inevitably suffer from extreme computational burdens as they need to train the same number of models as multiple target classes. Therefore, single-target attacks become impractical in attacking real classification systems that often have hundreds of categories. To pursue the extreme speed and storage, MAN trains only *one* model for 1000 target categories from ImageNet [9]. However, the excess of the target class severely degrades the targeted

transferability. C-GSP [67] improves the performance by designing a hierarchical partition mechanism to divide all classes into a feasible number of subsets and train generators for each subset. Nonetheless, C-GSP simply conditions the generator with one-hot vectors and only utilizes the classification information to train the generator, failing to use the semantic knowledge of the target class. To overcome the limitation, we propose to incorporate the text information provided by CLIP as significant guidance for the target class.

3 Method

In this section, we first introduce the preliminaries of targeted transferable attacks and present the basic paradigm of generative attack methods. Then, we elaborate on the proposed CGNC, which remarkably enhances multi-target black-box attacks. Finally, we detail the proposed masked fine-tuning technique that strengthens our method in single-class attacks.

3.1 Preliminary

We denote the white-box image classifier parameterized with θ as $f_\theta : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^{N\times H\times W}$ represents the image domain and $\mathcal{Y} \subset \mathbb{R}^L$ is the output confidence score of different classes (H, W, N, L being height, width, number of channels, and class number). Given a natural image $\boldsymbol{x} \in \mathcal{X}$ and the attacker's desired label $c_t \in \mathcal{C}$, the transferable targeted adversarial attacks attempt to craft the imperceptible perturbation $\boldsymbol{\delta}$ based on the accessible surrogate model f_θ to mislead another unseen victim model F_ϕ into predicting c_t, *i.e.*, $\arg\max_{i\in\mathcal{C}} F_\phi(\boldsymbol{x}+\boldsymbol{\delta})_i = c_t$. Concurrently, the l_∞ norm is employed to ensure the adversarial samples are indistinguishable from the original images by constraining the perturbation within the range ϵ by $\|\boldsymbol{\delta}\|_\infty \le \epsilon$.

To boost the fooling rate of targeted black-box attacks, single-target attacks [48, 49, 51] utilize powerful generative models to learn the target data distribution and achieve higher transferability. However, these methods consume great computation time and resources for multi-target scenarios, making them impractical for real-world attacks. C-GSP [67] effectively solves this issue by formulating the multi-target attacks as learning a class-conditional generator G_w with weights w, which is capable of generating perturbations for any target class. Given an unlabeled training dataset $\mathcal{X}_s$, the optimization objective is as follows:

$$\min_w \mathbb{E}_{\boldsymbol{x}_s\sim\mathcal{X}_s, c\sim\mathcal{C}}[\mathcal{L}\big(f_\theta(\boldsymbol{x}_s + G_w(\boldsymbol{x}_s, c)\big), c)], \text{ s.t. } \|G_w(\boldsymbol{x}_s, c)\|_\infty \le \epsilon, \tag{1}$$

where $\mathcal{L}(\cdot,\cdot)$ denotes the cross-entropy (CE) loss. By minimizing the loss of specified classes using various unlabeled images from $\mathcal{X}_s$, we optimize the parameters w of the generative model and finally obtain a targeted conditional generator that can generate perturbation for any given clean image $\boldsymbol{x}_t$ from the test dataset $\mathcal{X}_t$. Specifically, the adversary can simply specify a target label c and craft an corresponding adversarial example via $\boldsymbol{x}_{adv} = \boldsymbol{x}_t + G_w(\boldsymbol{x}_t, c)$.

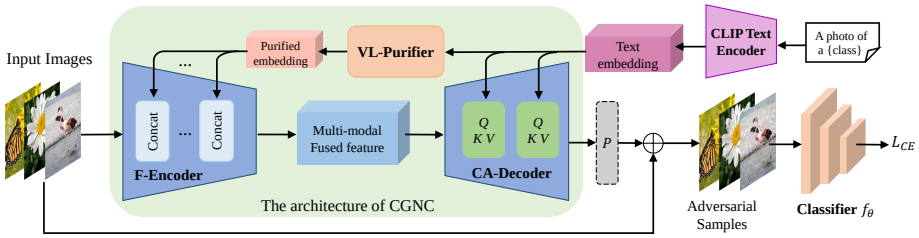

Fig. 2: An overview of our proposed architecture of CGNC. Equipped with the three exquisite modules VL-Purifier, F-Encoder, and CA-Decoder, the generator fully leverages the textual representations encoded by CLIP as auxiliary information about the target classes to better probe their data distribution and achieve superior attack effects.

However, current multi-target methods [27, 67] simply condition the generator with class labels and learn the target distribution only relying on the classification information of the surrogate model, thus not fully exploiting the specific characteristics of the target category. Inspired by the efficiency of vision-language (VL) learning [52, 59], we propose a novel generative network that leverages sufficient prior knowledge from the powerful CLIP model by incorporating textual-modality information to promote the target class modeling, which greatly enhances the multi-target transferable attacks.

3.2 CLIP-Guided Generative Network

The proposed generative model architecture is presented in Fig. 2. Specifically, CGNC is composed of a Vision-Language feature Purifier (VL-Purifier), a feature Fusion Encoder (F-Encoder), and a Cross-Attention based Decoder (CA-Decoder). We also provide the pseudocode of the training procedure in Algorithm 1. Next, we illustrate the design of each module as follows.

Vision-language feature purifier (VL-Purifier). To utilize CLIP to produce semantic embeddings, we first feed in CLIP's text encoder Φ with queries $\boldsymbol{t}_c$ that follow the handcrafted template "a photo of a {class}", which has shown effectiveness in many tasks [33, 52]. Since the obtained embeddings $\boldsymbol{e}_t \in \mathbb{R}^{B\times 512}$ (B being the batch size) in CLIP's vision-language space are generic representations of the target classes and are not yet tailored to our learning task, we refine them using the VL-Purifier, which is composed of several blocks consisting of a fully-connected layer and a spectral normalization layer. Through this module, we translate the encoded embeddings into more meaningful representations $\boldsymbol{e}_t^* \in \mathbb{R}^{B\times 16}$, thereby facilitating the subsequent step of feature fusion.

Feature fusion encoder (F-Encoder). This module aims to fuse the purified features $\boldsymbol{e}_t^*$ with the learned visual representations. Firstly, a batch of input images $\boldsymbol{x}_s$ is encoded to capture the visual concepts $\boldsymbol{h}_s \in \mathbb{R}^{B\times N'\times H'\times W'}$. Then, we expand the text embedding $\boldsymbol{e}_t^* \in \mathbb{R}^{B\times 16}$ into $\boldsymbol{e}_t^{*\prime} \in \mathbb{R}^{B\times 16\times H'\times W'}$, which are

Algorithm 1 Pseudocode of Training the CLIP-guided Generative Network

Require: $\mathcal{X}_s$: the training data; $\mathcal{C}$: the target label space; $\mathcal{T}$: the text prompts set; f_θ: the surrogate model; Φ: the CLIP's text encoder; N: the max iteration;
Ensure: the perturbation generator G_w;
for $i \leftarrow 0$ to N **do**
  Sample a batch of images $\boldsymbol{x}_s \sim \mathcal{X}_s$;
  Obtain $\boldsymbol{x}'_s$ by processing $\boldsymbol{x}_s$ with data augmentation;
  Sample a batch of target labels $c \sim \mathcal{C}$;
  Obtain the corresponding text prompts $\boldsymbol{t}_c$ from $\mathcal{T}$;
  Compute the text embedding $\boldsymbol{e}_t$ by feeding $\boldsymbol{t}_c$ into Φ;
  Obtain perturbed images by $\boldsymbol{x}_{adv} = \boldsymbol{x}_s + G_w(\boldsymbol{x}_s, \boldsymbol{e}_t)$, $\boldsymbol{x}'_{adv} = \boldsymbol{x}'_s + G_w(\boldsymbol{x}'_s, \boldsymbol{e}_t)$;
  Forward pass $\boldsymbol{x}_{adv}$, $\boldsymbol{x}'_{adv}$ to f_θ and compute the loss in Eq. (3);
  Perform gradient backpropagation and update the generator G_w;
end for
return the trained generator G_w

then integrated with the extracted visual concepts $\boldsymbol{h}_s$ through channel-wise concatenation to obtain the fused representations, *i.e.*, $\boldsymbol{m} \in \mathbb{R}^{B\times(N'+16)\times H'\times W'}$. Subsequently, $\boldsymbol{m}$ undergoes further downsampling, and the resultant features are again concatenated with the expanded embedding $\boldsymbol{e}^{*\prime}_t$. By repeating this operation several times, we effectively fuse the visual concepts of input images and the purified CLIP's embedding of the target classes. This mechanism fully exploits both the instance-level and class-level information from visual and textual modalities, thus encouraging the generation of perturbations with better semantic patterns and higher transferability.

Cross-Attention based Decoder(CA-Decoder). Given the multi-modal fused features from the previous module, the decoder attempts to translate them into perturbations of the target class. The network backbone is realized based on the decoder used in previous works [49, 67]. To fully explore the semantic priors brought by the CLIP model, we enhance the underlying backbone by introducing the cross-attention mechanism, which is proven to be effective for many multi-modal learning tasks. Specifically, we incorporate the textual embedding $\boldsymbol{e}_t$ from CLIP's latent space into our decoder via the cross-attention layer:

$$\begin{aligned} Q = \boldsymbol{z}_t W_q, K = \boldsymbol{e}_t W_k, V = \boldsymbol{e}_t W_v, \\ Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V, \end{aligned} \tag{2}$$

where $\boldsymbol{z}_t \in \mathbb{R}^{B\times d_\alpha}$ denotes the flattened intermediate features of the decoder, $W_q \in \mathbb{R}^{d_\alpha \times d}$, $W_k \in \mathbb{R}^{512\times d}$, $W_v \in \mathbb{R}^{512\times d}$ are learnable parameters. Similar to C-GSP, we post-process the output $\boldsymbol{o}$ via the $\tanh(\cdot)$ smooth projection to obtain the ℓ_∞ constrained perturbation with budge ϵ, *i.e.*, $\boldsymbol{\delta} = P(\boldsymbol{o}) = \epsilon \cdot \tanh(\boldsymbol{o})$.

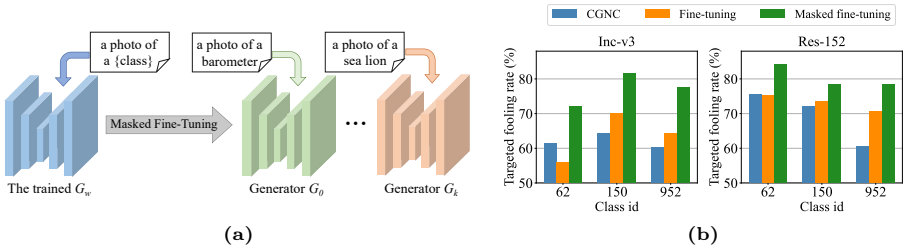

Fig. 3: Illustration of the proposed masked fine-tuning mechanism. (a) We fix the condition input with different text prompts to fine-tune the trained conditional generator G_w into multiple generators for single-target attacks. (b) The fooling rate of several target classes with Inc-v3 and Res-152 as substitute models respectively. The results indicate that direct fine-tuning yields inadequate results for certain classes due to overfitting. We efficiently resolve this issue via a patch-wise random mask operation.

Optimization objective. Based on the proposed CLIP-guided architecture, the optimization objective can be formulated as:

$$w^* \leftarrow \arg\min_{w} \mathcal{L}\Big(f_\theta\big(\boldsymbol{x}_s + G_w(\boldsymbol{x}_s, \Phi(\boldsymbol{t}_c))\big), c\Big), \tag{3}$$

where $\boldsymbol{t}_c$ represents the corresponding text prompts of the target classes c, $\mathcal{L}(\cdot,\cdot)$ denotes the cross-entropy loss. By encouraging the network to craft adversarial samples that can misguide the surrogate model's output toward the desired labels, the generator learns the data distribution of the target class and thus exhibits great generalizability in producing perturbation for any input data.

3.3 Masked Fine-Tuning Mechanism

In addition to the proposed CLIP-guided generator for multi-target attack scenarios, we also design a *single-target* variant for further improved performance. As depicted in Fig. 3a, we fix the conditioning input with the text description of a specific target class, and fine-tune the trained multi-target generator by resolving the objective function in Eq. (3) on the unlabeled training data $\mathcal{X}_s$. This strategy further enhances the attack effects by fine-tuning the conditional generator to specialize in a specific target class.

However, we encounter the overfitting problem that leads to limited improvement in success rates or even performance degradation for certain target classes. This problem is partly attributed to the fact that the generated adversarial perturbation sometimes heavily focuses on specific regions of the input image [62]. To alleviate this issue, we adopt a patch-wise random mask operation to post-process the adversarial perturbation produced by the generator, which brings a notable increase in the targeted fooling rate as shown in Fig. 3b. This technique improves the capability and flexibility of our method, as an adversary can perform fine-tuning to further augment the generator for single-target attacks, achieving a better trade-off between efficiency and performance compared to previous single-target methods (See Section 4.5 for detailed analysis).

4 Experiments

To validate the effectiveness of CGNC in boosting transferability, we conduct extensive experiments on various black-box models in a range of scenarios. Our code is availabel at https://github.com/ffhibnese/CGNC_Targeted_Adversarial_Attacks. Please see the Appendix for more experimental results.

4.1 Experimental Settings

Datasets. Following [19,67], we train the generator on the ImageNet training set [9] and evaluate the attack performance using ImageNet-NeurIPS (1k) dataset proposed by [50]. We also conduct experiments in a more realistic cross-domain scenario where we train the generator and generate perturbations on MS-COCO [41] or Comics [4] while evaluating on target classifier trained on ImageNet.

Victim Models. We use various victim models with different architectures. Specifically, the naturally trained models include Inception-v3 (Inc-v3) [57], Inception-v4 (Inc-v4) [55], Inception-ResNet-v2 (Inc-Res-v2) [55], ResNet-152 (Res-152) [29], DenseNet-121 (DN-121) [32], GoogleNet [56], and VGG-16 [54].

For further evaluation, we also analyze on the robust-trained models, including adv-Inception-v3 ($\text{Inc-v3}_{\text{adv}}$) [26], ens-adv-Inception-ResNet-v2 ($\text{IR-v2}_{\text{ens}}$) [28], and robust-trained ResNet-50 [24,31], dubbed as $\text{Res50}_{\text{SIN}}$ (trained on stylized ImageNet), Res50_{IN} (trained on the mixture of stylized and Nature ImageNet), $\text{Res50}_{\text{fine}}$ (Res50_{IN} plus further finetuning with an auxiliary dataset [24]), and $\text{Res50}_{\text{Aug}}$ (trained with the advanced data augmentation Augmix [31]).

Baseline Attacks. We reveal the superiority of the proposed CGNC in enhancing multi-target attacks by comparing our method with multiple competitive baselines, including MIM [11], DIM [65], SIM [40], DIM [12], Logit [72], SU [63], and the state-of-the-art (SOTA) multi-target generative attacks C-GSP [67] in Section 4.2. For SU attack [63], we choose to compare with its best version DTMI-Logit-SU. Besides, we also provide a comparison of the proposed single-target variant of CGNC with the existing single-target attack methods, *i.e.*, GAP [51], CD-AP [48], TTP [49], and DGTA-PI [19] in Section 4.5.

Implementation Details. Following [19,67], we employ Inc-v3 and Res-152 as surrogate models and adopt the perturbation budget ϵ 16/255. We conduct 10 epochs of generator training using a learning rate of 2e-4. During the masked fine-tuning, we maintain the learning rate and apply a mask ratio of 0.2 to fine-tune the text-conditional generator for an additional 5 epochs.

4.2 Multi-Target Transferability Evaluation

To align the experimental setup to former works [19,67], we first target 8 different classes in [71] and provide the average attack success rates (ASR) of the 8 target categories as evaluation metrics. Attack performance under larger numbers of target categories is presented in the next section.

Attacks against regularly trained models. We first perform attacks on normal models to evaluate the multi-target attack performance. By observing

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

Source	Method	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogleNet	VGG-16
Inc-v3	MIM	**99.90***	0.80	1.00	0.40	0.20	0.20	0.30
	TI-MIM	98.50*	0.50	0.50	0.30	0.20	0.40	0.40
	SI-MIM	99.80*	1.50	2.00	0.80	0.70	0.70	0.50
	DIM	95.60*	2.70	0.50	0.80	1.10	0.40	0.80
	TI-DIM	96.00*	1.10	1.20	0.50	0.50	0.50	0.80
	SI-DIM	90.20*	3.80	4.40	2.00	2.20	1.70	1.40
	Logit	99.60*	5.60	6.50	1.70	3.00	0.80	1.50
	SU	99.59*	5.80	7.00	3.35	3.50	2.00	3.94
	C-GSP	93.40*	46.58	36.74	41.60	46.40	40.00	45.00
	CGNC	96.03*	**59.43**	**48.06**	**42.48**	**62.98**	**51.33**	**52.54**
Res-152	MIM	0.50	0.40	0.60	**99.70***	0.30	0.30	0.20
	TI-MIM	0.30	0.30	0.30	96.50*	0.30	0.40	0.30
	SI-MIM	1.30	1.20	1.60	99.50*	1.00	1.40	0.70
	DIM	2.30	2.20	3.00	92.30*	0.20	0.80	0.70
	TI-DIM	0.80	0.70	1.00	90.60*	0.60	0.80	0.50
	SI-DIM	4.20	4.80	5.40	90.50*	4.20	3.60	2.00
	Logit	10.10	10.70	12.80	95.70*	12.70	3.70	9.20
	SU	12.36	11.31	16.16	95.08*	16.13	6.55	14.28
	C-GSP	37.70	33.33	20.28	93.20*	64.20	41.70	45.90
	CGNC	**53.39**	**51.53**	**34.24**	95.85*	**85.66**	**62.23**	**63.36**

the results in Table 1, we demonstrate that the proposed CGNC consistently achieves significant improvement compared with previous methods. Specifically, our method achieves an average improvement of 17.88% and 10.08% in ASR over the C-GSP [67] attack regarding Res-152 and Inc-v3 as surrogate models respectively, demonstrating the superiority of leveraging the rich prior knowledge from CLIP's text embedding. We also note that the iterative methods obtain nearly 100% while-box ASR while receiving poor performance on black-box models due to overfitting the classification boundaries of surrogate models.

Attacks under defense strategies. For a more thorough analysis and comparison, we then compare these attacks under several widely used defenses. Firstly, we consider attacking six robustly trained networks in Table 2. Generally, our method is still able to deceive the black-box classifiers into predicting the specified classes, significantly outperforming former multi-target attacks, *e.g.*, a 20.87 % increase of fooling rate from Res-152 to $\text{Res50}_{\text{Aug}}$.

Next, we evaluate the performance of different approaches on models with input preprocessing-based defenses, including a set of image smoothing mechanisms [10] and JPEG compression [14] algorithms. As shown in Table 3, although these defenses eliminate certain valid information in the adversarial samples, the CLIP-empowered CGNC still demonstrates excellent capability in boosting targeted transferability. Particularly when the substitute model is Res-152, CGNC achieves an average fooling rate of 71.18% and 80.34% on smoothing methods and JPEG compression, while C-GSP only reaches 49.05% and 56.16% respectively, verifying the stability and robustness of the proposed method.

Perturbation Visualization. We present visualization results in Fig. 4 to unveil the principle of our method. Upon careful inspection, we can see that the generated perturbations mainly focus on the semantic regions of the input

Table 2: Comparison of the proposed CGNC with the SOTA multi-target attacks against models with robust training mechanism on ImageNet NeurIPS validation set.

Source	Method	Inc-v3$_{adv}$	IR-v2$_{ens}$	Res50$_{SIN}$	Res50$_{IN}$	Res50$_{fine}$	Res50$_{Aug}$
Inc-v3	MIM	0.16	0.10	0.20	0.27	0.44	0.19
	TI-MIM	0.21	0.19	0.33	0.49	0.68	0.31
	SI-MIM	0.13	0.19	0.26	0.43	0.63	0.29
	DIM	0.11	0.09	0.16	0.33	0.39	0.19
	TI-DIM	0.15	0.13	0.16	0.21	0.33	0.14
	SI-DIM	0.19	0.21	0.43	0.71	0.84	0.46
	Logit	0.30	0.30	0.70	1.23	3.14	0.86
	SU	0.49	0.41	0.84	1.75	3.55	1.04
	C-GSP	20.41	18.04	6.96	33.76	44.56	21.95
	CGNC	**24.36**	**22.54**	**8.85**	**40.83**	**52.18**	**22.85**
Res-152	MIM	0.19	0.15	0.28	1.58	2.75	0.78
	TI-MIM	0.61	0.73	0.50	2.51	4.75	1.76
	SI-MIM	0.24	0.24	0.39	0.66	0.84	0.36
	DIM	0.63	0.37	0.94	8.50	14.22	3.77
	TI-DIM	0.23	0.30	0.28	0.76	1.49	0.49
	SI-DIM	0.71	0.71	0.75	2.73	3.89	1.37
	Logit	1.15	1.18	1.65	6.70	15.46	5.93
	SU	2.12	1.20	1.95	7.53	21.14	6.95
	C-GSP	14.60	16.01	16.84	60.30	65.51	42.88
	CGNC	**22.21**	**26.71**	**29.83**	**79.80**	**84.05**	**63.75**

Table 3: Comparison of our method with C-GSP under different defense strategies. Q indicates the quality factor in JPEG compression. Here the target model is DN-121 and results for more victim models are in the Appendix.

Source	Method	Smoothing			JPEG compression				
		Gaussian	Medium	Average	Q=70	Q=75	Q=80	Q=85	Q=90
Inc-v3	C-GSP	36.59	46.91	39.86	49.41	50.55	51.95	52.84	53.39
	CGNC	**43.03**	**55.35**	**45.94**	**58.35**	**59.25**	**60.28**	**61.29**	**61.94**
Res-152	C-GSP	43.21	56.38	47.55	53.03	54.54	56.13	57.75	59.36
	CGNC	**64.44**	**79.84**	**69.25**	**77.38**	**78.81**	**80.54**	**82.05**	**82.94**

images and contain sufficient semantic patterns specific to the target category. For instance, when the condition is *a photo of a sea lion*, the resulting perturbations indeed carry rich patterns closely resembling this marine animal. We also observe that the pattern changes in accordance with the text prompts, which validates our conditioning mechanism of the CLIP-encoded textual embedding.

4.3 Evaluation on More Scenarios

Cross-Domain scenarios. Next, we explore the more realistic cross-domain scenarios [48, 49] where attackers do not know anything about the data distribution of the training set used by the black-box classifier. Based on this setting, attackers train the generator and generate adversarial perturbations using an auxiliary dataset that follows a different probability distribution from that of the target model. Specifically, we satisfy the cross-domain experimental setting using MS-COCO [41] and Comics [4] datasets respectively. MS-COCO is a large-scale image dataset widely used for object detection and semantic segmentation,

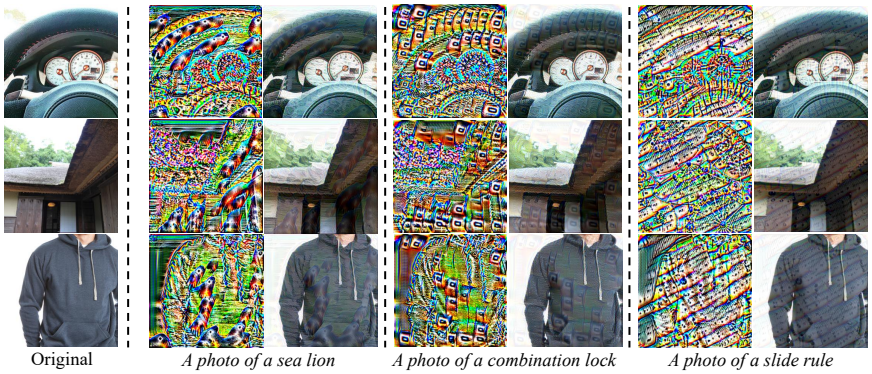

Fig. 4: Visualization results of different input images for different targets. For each text prompt of the target class, the left column shows the perturbation generated by our CGNC while the right column displays the corresponding adversarial examples.

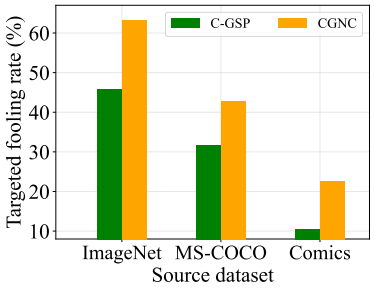

Fig. 5: Fooling rates (from Res-152 to VGG-16) in attacking 8 target classes on cross-domain scenarios. We also provide the results on ImageNet as a comparison.

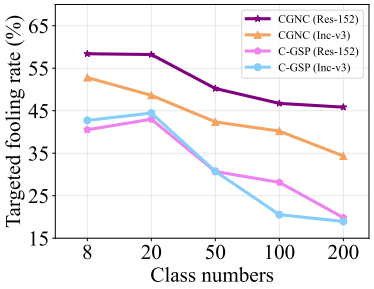

Fig. 6: Fooling rates of our CGNC with C-GSP on larger numbers of attacked classes regarding Res-152 and Inc-v3 as surrogate models.

while Comic is composed of a large number of comic images that can be regarded as a stylized version of ImageNet.

To analyze the cross-domain transferability, we randomly select 1000 images from the source dataset to craft adversarial samples using the trained generator. Fig. 5 shows the results of the cross-domain attacks transferring from three different source datasets to ImageNet. In general, the more challenging cross-domain datasets lead to varying degrees of reduction in the targeted fooling rate due to the domain gap, especially on the Comics dataset which differs greatly from the ImageNet distribution. Benefiting from the text guidance on the target category, CGNC still attains decent attack performance and remarkably outperforms C-GSP. This demonstrates that our method achieves better cross-domain transferability and is partly independent of the training dataset.

In addition, since the SOTA single-target attacks [19, 49, 62] require samples of the target class for loss computation, they are not applicable in the cross-

Table 4: Ablation study of CGNC and its two variants on ImageNet NeurIPS validation set for 8 target classes. The substitute model is Res-152.

Method	VGG-16	GoogleNet	Inc-v3	Res-152	DN-201
CGNA-CA-t	56.55	51.09	47.44	92.81	74.65
CGNA-CA	56.64	54.29	49.73	93.34	75.99
CGNC	63.36	62.23	53.39	95.85	82.69

domain scenarios where the source dataset lacks images of the target category. This represents an additional advantage of our method over single-target attacks.

Larger Numbers of Target Classes. We then increase the number of target categories to verify the effectiveness of our method when handling plenty of classes. Specifically, we condition CGNC with more text inputs corresponding to the increased number of target categories. The performance is evaluated across the aforementioned six black-box models.

As mentioned before, single-target attacks become impractical for real-world classification systems with hundreds or thousands of target categories [27]. In comparison, our method effectively solves this issue and achieves great improvements in large numbers of target classes over C-GSP as in Fig. 6, *e.g.*, 14.71% increase for 200 target classes proxy on the Res-152. This again reveals the significance of the fully exploited textual guidance from the CLIP model. Additionally, our method exhibits greater robustness to the varying numbers of target classes, *e.g.*, the performance of our network on Inc-v3 demonstrates a smoother decrease compared to that of C-GSP. This further highlights the superiority of our method in reducing computation costs. For instance, when launching 1000-class targeted attacks using Inc-v3 as the surrogate model, our method requires training only 5 CGNCs with 200 conditions, yet it can achieve comparable performance to 20 generators with 50 conditions trained using C-GSP.

4.4 Ablation Study

We conduct ablation experiments on the ImageNet NeurIPS dataset to study the effect of the proposed techniques. Specifically, we introduce two variants of CGNC. CGNC-CA removes the cross-attention module from the original network, and CGNC-CA-t further modifies the conditioning mechanism by replacing the CLIP's text embedding with one-hot labels.

From the results in Table 4, we can find that each aforementioned technique can further improve the attack success rates. Moreover, the remarkable improvement from CGNC-CA to CGNC also confirms that these cross-attention modules in the CA-Decoder can make better use of the text guidance provided by the CLIP model to enhance the targeted transferability of crafted perturbations.

4.5 Comparison with Single-Target Attacks

Next, we compare the single-target variants of the proposed CGNC with multiple state-of-the-art single-target attacks. To obtain our single-target generators, we

Table 5: Comparison of our method with the SOTA single-target attacks. † denotes the single-target variant of our CGNC obtained through the masked fine-tuning technique. * represents the white-box attacks.

Source	Method	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogleNet	VGG-16
Inc-v3	GAP	86.90*	45.06	34.48	34.48	41.74	26.89	34.34
	CD-AP	94.20*	57.60	60.10	37.10	41.60	32.30	41.70
	TTP	91.37*	46.04	39.37	16.40	33.47	25.80	25.73
	DGTA-PI	94.63*	67.95	55.03	50.50	47.38	47.67	48.11
	CGNC†	**98.84***	**74.76**	**64.48**	**62.00**	**78.94**	**69.06**	**70.74**
Res-152	GAP	30.99	31.43	20.48	84.86*	58.35	29.89	39.70
	CD-AP	33.30	43.70	42.70	96.60*	53.80	36.60	34.10
	TTP	62.03	49.20	38.70	95.12*	82.96	65.09	62.82
	DGTA-PI	66.83	53.62	**47.61**	96.48*	86.61	68.29	**69.58**
	CGNC†	**68.86**	**69.45**	45.71	**98.61***	**91.14**	**69.83**	68.05

conduct the proposed masked fine-tuning to the trained conditional generator eight times using eight different text prompts of the target classes. Quantitative results in Table 5 indicate that our single-target enhanced generators outperform the competing single-target attacks in most cases. Particularly on the surrogate model of Inc-v3, CGNC† achieves a notable increase of 15.36% in the average black-box fooling rate compared to previous methods, demonstrating the significant effectiveness of the proposed masked fine-tuning. The MFT technique further enhances the single-target performance, thereby improving the adaptability and flexibility of our approach in scenarios with fewer target classes.

Note that these single-target methods require training 8 generators from scratch for 8 different target classes. In contrast, our method only needs to train a single multi-target generator and perform fine-tuning 8 times, each time with just a few epochs. Based on the typical experimental setup used in the SOTA single-target methods [19,49], our strategy can diminish over 100 training epochs when targeting 8 classes, thus substantially mitigating the computational burden. More experiments of single-target attacks against adversarially robust models and input preprocessing defenses are shown in the Appendix.

5 Conclusion

In this paper, we design a novel generative network CGNC, which improves multi-target transferable adversarial attacks by fully utilizing the rich prior within the CLIP as auxiliary semantic knowledge about the target category. To better incorporate the prior information, we introduce the cross-attention modules and efficiently condition the generator with CLIP's text embeddings. Through extensive experiments, we demonstrate the effectiveness of the proposed CGNC on multiple black-box target models in a variety of scenarios. Moreover, we propose a masked fine-tuning technique to further enhance the CGNC in attacking a single class, which outperforms previous single-target methods in both efficiency and effectiveness. We hope that the proposed method can serve as a reliable tool for evaluating the model robustness under black-box setups, promoting further research on the vulnerability and robustness of DNNs.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grant 62171248, 62301189, Guangdong Basic and Applied Basic Research Foundation under grant 2021A1515110066, the PCNL KEY project (PCL2021A07), and Shenzhen Science and Technology Program under Grant JCYJ20220818101012025, RCBS20221008093124061, GXWD20220811172936001.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. Ieee Access **6**, 14410–14430 (2018)
2. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European conference on computer vision. pp. 484–501. Springer (2020)
3. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: International conference on machine learning. pp. 284–293. PMLR (2018)
4. BircanoAYlu, C.: `https://www.kaggle.com/datasets/cenkbircanoglu/comic-books-classification`. Kaggle, 2017
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
6. Che, Z., Borji, A., Zhai, G., Ling, S., Li, J., Min, X., Guo, G., Le Callet, P.: Smgea: A new ensemble adversarial attack powered by long-term gradient memories. IEEE Transactions on Neural Networks and Learning Systems **33**(3), 1051–1065 (2020)
7. Chefer, H., Benaim, S., Paiss, R., Wolf, L.: Image-based clip-guided essence transfer. In: European Conference on Computer Vision. pp. 695–711. Springer (2022)
8. Chen, B., Feng, Y., Dai, T., Bai, J., Jiang, Y., Xia, S.T., Wang, X.: Adversarial examples generation for deep product quantization networks on image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 1388–1404 (2022)
9. 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)
10. Ding, G.W., Wang, L., Jin, X.: Advertorch v0. 1: An adversarial robustness toolbox based on pytorch. arXiv preprint arXiv:1902.07623 (2019)
11. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
12. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4312–4321 (2019)
13. Du, J., Zhang, H., Zhou, J.T., Yang, Y., Feng, J.: Query-efficient meta attack to deep neural networks. arXiv preprint arXiv:1906.02398 (2019)
14. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853 (2016)
15. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1625–1634 (2018)

16. Fang, H., Chen, B., Wang, X., Wang, Z., Xia, S.T.: Gifd: A generative gradient inversion method with feature domain optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4967–4976 (2023)
17. Fang, H., Kong, J., Yu, W., Chen, B., Li, J., Xia, S., Xu, K.: One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. arXiv preprint arXiv:2406.05491 (2024)
18. Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., Chen, B., Wang, X., Xia, S.T.: Privacy leakage on dnns: A survey of model inversion attacks and defenses. arXiv preprint arXiv:2402.04013 (2024)
19. 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)
20. Gao, K., Bai, Y., Bai, J., Yang, Y., Xia, S.T.: Adversarial robustness for visual grounding of multimodal large language models. In: ICLR Workshop (2024)
21. Gao, K., Bai, Y., Gu, J., Xia, S.T., Torr, P., Li, Z., Liu, W.: Inducing high energy-latency of large vision-language models with verbose images. In: ICLR (2024)
22. Gao, K., Gu, J., Bai, Y., Xia, S.T., Torr, P., Liu, W., Li, Z.: Energy-latency manipulation of multi-modal large language models via verbose samples. arXiv preprint arXiv:2404.16557 (2024)
23. Gao, L., Cheng, Y., Zhang, Q., Xu, X., Song, J.: Feature space targeted attacks by statistic alignment. arXiv preprint arXiv:2105.11645 (2021)
24. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
26. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
27. Han, J., Dong, X., Zhang, R., Chen, D., Zhang, W., Yu, N., Luo, P., Wang, X.: Once a man: Towards multi-target attack via learning multi-target adversarial network once. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5158–5167 (2019)
28. Hang, J., Han, K., Chen, H., Li, Y.: Ensemble adversarial black-box attacks against deep learning systems. Pattern Recognition **101**, 107184 (2020)
29. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 630–645. Springer (2016)
30. He, Z., Wang, W., Dong, J., Tan, T.: Revisiting ensemble adversarial attack. Signal Processing: Image Communication **107**, 116747 (2022)
31. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)
32. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
33. 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)

34. Inkawhich, N., Liang, K., Wang, B., Inkawhich, M., Carin, L., Chen, Y.: Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. Advances in Neural Information Processing Systems **33**, 20791–20801 (2020)
35. Inkawhich, N., Wen, W., Li, H.H., Chen, Y.: Feature space perturbations yield more transferable adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7066–7074 (2019)
36. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: European Conference on Computer Vision. pp. 105–124. Springer (2022)
37. Kong, Z., Guo, J., Li, A., Liu, C.: Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14254–14263 (2020)
38. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
39. Li, Q., Guo, Y., Chen, H.: Yet another intermediate-level attack. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 241–257. Springer (2020)
40. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. arXiv preprint arXiv:1908.06281 (2019)
41. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
42. Liu, A., Wang, J., Liu, X., Cao, B., Zhang, C., Yu, H.: Bias-based universal adversarial patch attack for automatic check-out. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 395–410. Springer (2020)
43. Lu, Y., Jia, Y., Wang, J., Li, B., Chai, W., Carin, L., Velipasalar, S.: Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 940–949 (2020)
44. Luo, J., Bai, T., Zhao, J.: Generating adversarial yet inconspicuous patches with a single image (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 15837–15838 (2021)
45. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
46. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
47. Mopuri, K.R., Uppala, P.K., Babu, R.V.: Ask, acquire, and attack: Data-free uap generation using class impressions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 19–34 (2018)
48. 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)
49. 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)

50. NeurIPS: `https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack/data`. Kaggle, 2017
51. 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)
52. 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)
53. Sarkar, S.K., Oshiba, K., Giebisch, D., Singer, Y.: Robust classification of financial risk. arXiv preprint arXiv:1811.11079 (2018)
54. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
55. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
56. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
57. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
58. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
59. Tao, M., Bao, B.K., Tang, H., Xu, C.: Galip: Generative adversarial clips for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14214–14223 (2023)
60. Wang, X., He, K.: Enhancing the transferability of adversarial attacks through variance tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1924–1933 (2021)
61. Wang, X., He, K., Hopcroft, J.E.: At-gan: A generative attack model for adversarial transferring on generative adversarial nets. arXiv preprint arXiv:1904.07793 **3**(4), 3 (2019)
62. Wang, Z., Yang, H., Feng, Y., Sun, P., Guo, H., Zhang, Z., Ren, K.: Towards transferable targeted adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20534–20543 (2023)
63. 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)
64. Xiao, C., Li, B., Zhu, J.Y., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610 (2018)
65. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2730–2739 (2019)
66. Xiong, Y., Lin, J., Zhang, M., Hopcroft, J.E., He, K.: Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14983–14992 (2022)

67. 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)
68. Yu, W., Fang, H., Chen, B., Sui, X., Chen, C., Wu, H., Xia, S.T., Xu, K.: Gi-nas: Boosting gradient inversion attacks through adaptive neural architecture search. arXiv preprint arXiv:2405.20725 (2024)
69. Zha, Y., Ji, H., Li, J., Li, R., Dai, T., Chen, B., Wang, Z., Xia, S.T.: Towards compact 3d representations via point feature enhancement masked autoencoders. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6962–6970 (2024)
70. Zha, Y., Wang, J., Dai, T., Chen, B., Wang, Z., Xia, S.T.: Instance-aware dynamic prompt tuning for pre-trained point cloud models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14161–14170 (2023)
71. Zhang, C., Benz, P., Imtiaz, T., Kweon, I.S.: Understanding adversarial examples from the mutual influence of images and perturbations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14521–14530 (2020)
72. 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)