Rethinking Target Label Conditioning in Adversarial Attacks: A 2D Tensor-Guided Generative Approach

Hangyu Liu[1,2], **Bo Peng**[3], **Pengxiang Ding**[1,2], **Donglin Wang**[2*]
[1]Zhejiang University
[2]Westlake University
[3]Beijing University of Posts and Telecommunications
12563051@zju.edu.cn, bonovice@bupt.edu.cn, dingpengxiang@westlake.edu.cn, wangdonglin@westlake.edu.cn

Abstract

Compared to single-target adversarial attacks, multi-target attacks have garnered significant attention due to their ability to generate adversarial images for multiple target classes simultaneously. However, existing generative approaches for multi-target attacks primarily encode target labels into one-dimensional tensors, leading to a loss of fine-grained visual information and overfitting to model-specific features during noise generation. To address this gap, we first identify and validate that the semantic feature quality and quantity are critical factors affecting the transferability of targeted attacks: 1) Feature quality refers to the structural and detailed completeness of the implanted target features, as deficiencies may result in the loss of key discriminative information; 2) Feature quantity refers to the spatial sufficiency of the implanted target features, as inadequacy limits the victim model's attention to this feature. Based on these findings, we propose the 2D Tensor-Guided Adversarial Fusion (TGAF) framework, which leverages the powerful generative capabilities of diffusion models to encode target labels into two-dimensional semantic tensors for guiding adversarial noise generation. Additionally, we design a novel masking strategy tailored for the training process, ensuring that parts of the generated noise retain complete semantic information about the target class. Extensive experiments demonstrate that TGAF consistently surpasses state-of-the-art methods across various settings.

Code — https://github.com/TemenosMistral/TGAF

1 Introduction

With the rapid advancement of deep neural networks (DNNs), artificial intelligence has achieved significant progress in fields such as medical diagnostics (Ma et al. 2021), image classification (He et al. 2016a), and autonomous driving (Kong et al. 2020). However, the adversarial vulnerability of deep learning models has emerged as a major challenge affecting their reliable application in security-sensitive scenarios.

Adversarial attacks manipulate input data by introducing subtle and imperceptible perturbations, misleading model predictions. These attacks can be categorized into two main types: untargeted attacks, which aim to cause misclassification without specifying a particular category, and targeted attacks, which force models to output labels chosen by the attacker. Targeted attacks are not only more challenging but also more threatening as they enable specific malicious control, such as deceiving an autonomous vehicle into interpreting a stop sign as a speed limit sign.

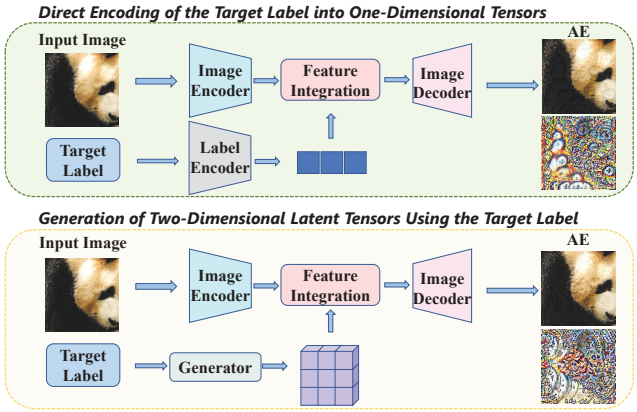

Figure 1: **Comparison of multi-target approaches.** Previous methods (top row) use 1D label encoding to guide noise generation for adversarial examples (AE). However, this often loses fine-grained details because images are 2D, potentially leading to overfitting. Our method (bottom row) generates 2D latent representations from target labels, better preserving structural information.

In real-world applications, attackers typically do not fully understand the architecture or parameters of the target model, making transferability a crucial property for effective adversarial attacks. Transferability refers to the ability of adversarial examples crafted on surrogate models to successfully deceive unknown black-box models (Wang and He 2021). While untargeted attacks have demonstrated strong transferability, targeted attacks still suffer from low success rates due to their dependency on overfitting the decision boundaries of surrogate models.

Existing targeted attack methods primarily fall into two categories: instance-specific approaches (Dong et al. 2018; Gao et al. 2021) and instance-agnostic approaches (Feng

*Corresponding author.

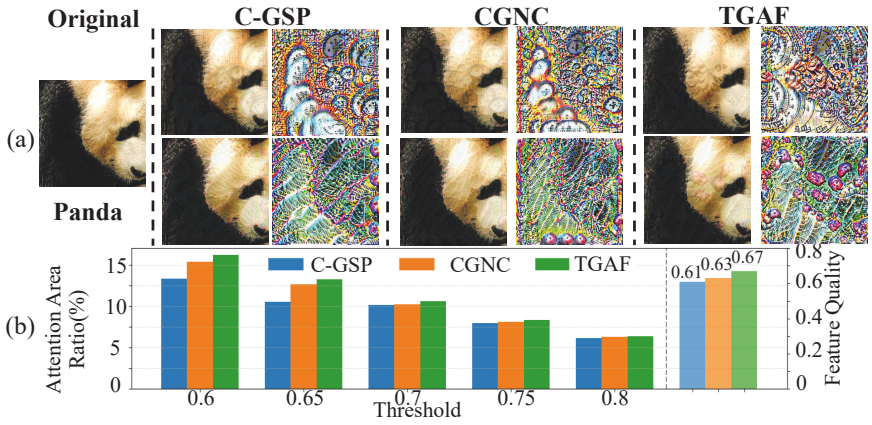

Figure 2: **(a) Visualization comparison of C-GSP, CGNC, and TGAF.** Each row displays an adversarial example and its corresponding perturbation map for a distinct target label: "barometer" (first row) and "fig" (second row). The surrogate model used is Inc-v3. TGAF demonstrably surpasses C-GSP and CGNC by more effectively capturing both the target semantic details (e.g., the barometer's pointer) and the target semantic quantity (e.g., the number of figs). **(b) Quantitative analysis of feature quantity and feature quality.** Feature quantity is measured by the percentage of high-attention area analyzed via Grad-CAM. Feature quality is measured by the cosine similarity between the perturbation's feature vector and the average feature vector of real target class images. Experiments are conducted on 1000 images.

et al. 2023; Kong et al. 2020). Instance-specific methods (Dong et al. 2018, 2019) optimize adversarial perturbations for individual samples in an iterative manner, often leading to inefficiency and susceptibility to overfitting. On the other hand, instance-agnostic approaches focus on learning universal perturbations (Moosavi-Dezfooli et al. 2017; Li et al. 2022) or generators (Naseer et al. 2019) based on data distribution to enhance attack generality. Recent generative techniques, such as C-GSP and CGNC, have significantly improved efficiency by training conditional generators for multiple target classes compared to single-target methods. However, these methods simply use 1D tensors to guide generation and fail to thoroughly investigate the key factors affecting attack transferability.

To address this gap, we observe that adversarial noise generated for targeted attacks functions similarly to "implanting" semantic features of the target class into source images. Building on this insight, we define two critical factors that influence attack transferability: the quality and quantity of target-class semantic features in the generated noise. Through experimental analysis (Fig. 3), we validate our hypothesis. Furthermore, we examine the adversarial noise produced by C-GSP and CGNC and identify issues related to incomplete semantic feature implantation and insufficient semantic representation, as shown in Fig. 2. We hypothesize that these shortcomings arise because these methods represent target labels as 1D tensors (e.g., one-hot encoding or CLIP embeddings), disregarding spatial and structural information, and tend to overfit to the regions where noise mapping is easiest during training. Consequently, the generated adversarial noise lacks essential target-class features and sufficient semantic information, thereby limiting its transferability.

To overcome these limitations, we propose a novel generative framework termed 2D-Tensor-Guided Adversarial Fusion (TGAF). The core innovation of our approach lies in leveraging 2D spatial information of target labels to guide adversarial noise generation. This enables the retention of fixed low-level semantic information during the generation process, rather than relying solely on the decision boundaries of surrogate models. Additionally, we introduce a carefully designed random masking strategy tailored for training, ensuring that parts of the generated noise still contain complete semantic information of the target class. Our contributions can be summarized as follows:

- We first systematically analyze adversarial noise generation from the perspective of semantic feature quality and quantity, uncovering their impact on transferability.
- We propose TGAF, an innovative approach that combines 2D target representation and random masking strategies to enhance adversarial transferability.
- Our method significantly outperforms previous attack methods in terms of targeted transferability across various experimental settings.

2 Related Work

2.1 Transferable Targeted Attacks

Transferable targeted adversarial attacks aim to fool multiple models into misclassifying adversarial examples into a specific target class. To enhance attack transferability, input transformation-based methods (Byun et al. 2022; Wei et al. 2023) diversify input representations through techniques such as geometric transformations and local mixup. Advanced objective-based approaches (Weng et al. 2023; Byun et al. 2023) refine loss functions or optimize specific outputs. Generation-based methods (Naseer et al. 2021; Zhao et al. 2023) employ generative models to produce adversarial examples that closely resemble the target class characteristics

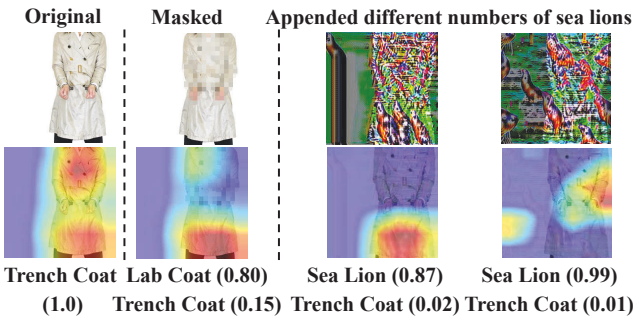

Figure 3: **Demonstrating the impact of feature implantation.** The heatmaps are generated on Res-152 using Grad-CAM. Left: The original image and heatmap. Middle: The masked image and heatmap (Masked trench coat buttons). Right: The perturbation and heatmap. Results confirm that 1) insufficient feature details reduce accuracy, while 2) more features have a higher target probability.

and ensemble-based techniques (Wu et al. 2024) focus on efficiently integrating multiple models or strategies, leveraging approaches like self-distillation and weight scaling.

2.2 Defense Methods

To mitigate adversarial attacks, various defense strategies have been developed. Among them, Adversarial Training (AT) (Goodfellow, Shlens, and Szegedy 2014) is one of the most effective, improving robustness by incorporating adversarial examples during training. Preprocessing-based defenses (Dziugaite, Ghahramani, and Roy 2016; Xu, Evans, and Qi 2017) instead remove adversarial perturbations before inference, while denoising techniques like HGD (Liao et al. 2018) use autoencoders guided by high-level features to purify inputs. Diffusion purification (Naseer et al. 2020; Nie et al. 2022) further enhances defense by leveraging diffusion models to reconstruct clean images.

3 Methodology

In this section, we first introduce the prerequisite knowledge and our motivation. Then, we provide a detailed explanation of our TGAF method.

3.1 Preliminaries

Let $f_\Phi : \mathcal{X} \to \mathcal{Y}$ denote a white-box image classifier parameterized by Φ, where $\mathcal{X} \in \mathbb{R}^{C\times H\times W}$ represents the image domain and $\mathcal{Y} \in \mathbb{R}^L$ denotes the output confidence scores over L classes. Given a natural image $x \in \mathcal{X}$ and a target class $c_t \in \mathcal{C}$, the goal of transferable targeted adversarial attacks is to craft an imperceptible perturbation δ such that the adversarial example $x_{adv} = x + \delta$ misleads both the surrogate model f_Φ and unknown black-box victim models F_Φ into predicting c_t. This objective can be formalized as follows:

$$\arg\max f_\Phi(x_{adv}) = c_t, \quad \text{with} \quad \|\delta\|_\infty \le \epsilon, \tag{1}$$

where ϵ constrains the perturbation magnitude under the l_∞-norm.

To achieve this objective, numerous instance-specific methods have been proposed. However, these approaches often require a large number of iterations per attack, resulting in high computational cost and low efficiency. To address this limitation, several multi-target methods have been developed, primarily based on a generative framework.

Existing generative approaches train a conditional perturbation generator G_θ to map input pairs (x, c_t) to targeted perturbations, i.e., $\delta = G_\theta(x, c_t)$. The optimization objective is formulated as:

$$\begin{aligned} \min_\theta \ \mathbb{E}_{x\sim\mathcal{X},\, c_t\sim\mathcal{C}}\Big[\mathcal{L}\big(f_\Phi\big(x + G_\theta(x,c_t)\big), c_t\big)\Big], \\ \text{s.t.} \quad \big\|G_\theta(x,c_t)\big\|_\infty \le \varepsilon. \end{aligned} \tag{2}$$

Here, $\mathcal{X}$ represents an unlabeled training set and $\mathcal{L}(\cdot,\cdot)$ is typically the cross-entropy loss.

3.2 Motivation

Although recent advances in targeted adversarial attacks have been made, the transferability of generated adversarial samples remains unsatisfactory. In non-targeted adversarial attacks, numerous studies (Wang et al. 2024; Liu et al. 2025) have highlighted that different models focusing on distinct object regions impede the generalization of adversarial samples. This observation naturally leads us to investigate:

What primarily hinders the generalizability of adversarial samples in targeted attacks?

Existing studies (Yang et al. 2022; Fang et al. 2024) suggest that targeted attacks achieve their objective by "implanting" target-class semantic features into source images. Based on this, we hypothesize that the nature of these implanted features is the key factor influencing attack success. Specifically, we examine: 1) **Feature quality**: The integrity and detail preservation of implanted features. 2) **Feature quantity**: The number of target-class features introduced into the source image. As shown in Fig. 3, our experiments reveal two critical phenomena: 1) When essential details (e.g., buttons on a trench coat) are masked, models exhibit high-confidence misclassifications (e.g., predicting "lab coat" instead of the true class). 2) Increasing the number of implanted target features (e.g., sea lion patches) raises the likelihood of target-class predictions. These two factors jointly affect adversarial sample generalizability. Since different models focus on disparate object regions, feature omissions may prevent certain models from capturing their critical discriminative features, while insufficient feature quantity further constrains attack effectiveness.

Based on our previous findings, we analyzed the adversarial noise generated by different multi-target approaches. As shown in Fig. 2 (a), we found that: 1) for the barometer class, the adversarial noise produced by C-GSP and CGNC methods lacks essential structural features including the pointer and scale markings; 2) for the fig class, these methods generate noise that primarily concentrates at image boundaries. These findings prompt a novel question:

How can we create adversarial noise that incorporates a greater quantity and higher completeness of target features?

To address this issue, we first conducted an analysis of existing approaches. Fig. 1 shows that current methods universally encode target labels as 1D tensors for processing.

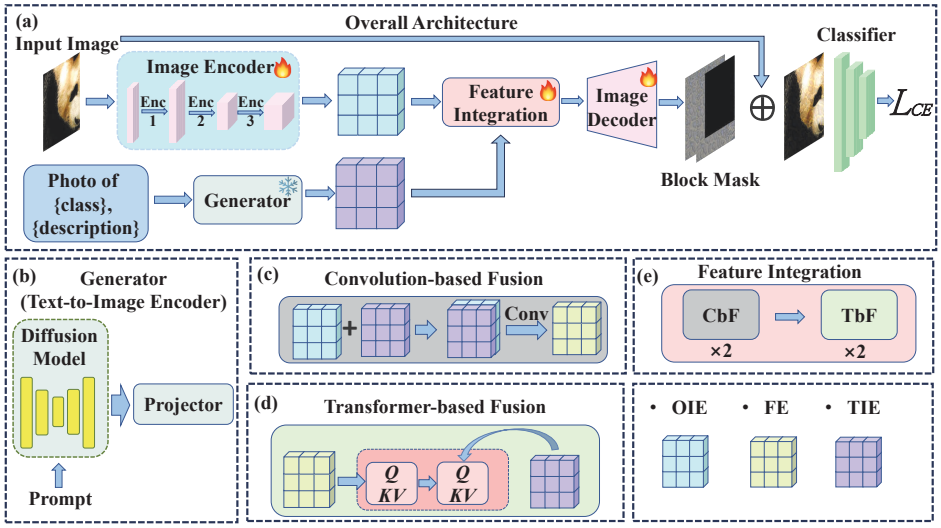

Figure 4: **The framework of 2D Target-Guided Adversarial Fusion (TGAF).** (a) Overall architecture; (b) Text-to-Image Encoder; (c) Convolution-based Fusion (CbF) module; (d) Transformer-based Fusion (TbF) module; (e) Feature Integration Module. Abbreviations: OIE (Original Image Embeddings), FE (Fusion Embeddings), TIE (Target Image Embeddings).

However, this 1D encoding scheme fundamentally differs from the 2D nature of images, resulting in the loss of low-level visual information. Such information deficiency introduces bias during training: the models tend to learn specific feature representations from surrogate models, consequently leading to overfitting.

Regarding the issue of insufficient feature injection, we developed an innovative masked training mechanism tailored for training inspired by CGNC's fine-tuning approach. This mechanism randomly masks the model's output noise during training, ensuring that partial noise regions retain complete feature information.

3.3 2D Target-Guided Adversarial Fusion

The architecture of TGAF method is illustrated in Fig. 4. Specifically, TGAF comprises four key components: an Image Encoder $\mathcal{E}$, a Text-to-Image Encoder $\mathcal{G}$, a Feature Integration Module $\mathcal{F}$, and an Image Decoder $\mathcal{D}$. We next detail the design of each component.

Image Encoder module. Given an input image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$, our encoder $\mathcal{E}$ extracts multi-scale features through three convolutional layers, gradually abstracting and downsampling the features. Specifically, we transform the image features from $C \times H \times W$ to $C' \times \frac{H}{4} \times \frac{W}{4}$. Let $\mathbf{x} \in \mathbb{R}^{C'\times \frac{H}{4}\times \frac{W}{4}}$ denote the encoded image representation.

Text-to-Image Encoder Module. To obtain a 2D representation of the target class, we leverage the capabilities of diffusion models, employing the encoder and denoising Unet parts of Stable-Diffusion-2 (Rombach et al. 2022) to process the input target labels, resulting in low-dimensional latent vectors of size $B \times 4 \times 64 \times 64$. To align with the features of the original image, we use a convolutional layer and an average pooling to project the latent vectors from $4 \times 64 \times 64$ to a latent representation $\mathbf{z}_c \in \mathbb{R}^{4\times \frac{H}{4}\times \frac{W}{4}}$.

Feature Integration Module. The goal of this module is to effectively fuse the original image representation $\mathbf{x}$ with the target-conditioned representation $\mathbf{z}_c$. We propose two fusion strategies:

1. Convolution-based Fusion (CbF): We employ a learnable 1×1 convolution to adaptively reduce channel dimensions and learn local feature interactions.

$$\mathbf{f}_c = \text{Conv}_{1\times 1}(\mathbf{x} \parallel \mathbf{z}_c), \tag{3}$$

where $\text{Conv}_{1\times 1}$ projects the concatenated features back to the original channel dimension C' of $\mathbf{x}$. Let x_c denote the output of convolution-based fusion.

2. Transformer-based Fusion (TbF): We leverage attention mechanisms to capture complex global spatial-channel dependencies and model long-range interactions between features. The TbF module comprises three sequential computational stages:

First, we transform the condition features z_c through a 1×1 convolutional layer to align the channel of x_c:

$$\mathbf{z}_t = \text{Conv}_{1\times 1}(\mathbf{z}_c). \tag{4}$$

Subsequently, we apply channel attention (Hu, Shen, and Sun 2018) to recalibrate feature importance:

$$\text{CHA}(\mathbf{x}) = \sigma\left(W_2 \cdot \text{ReLU}(W_1 \cdot \text{AvgPool}(\mathbf{x}))\right), \tag{5}$$

$$\mathbf{x}_{ca} = \mathbf{x}_c \odot \text{CHA}(\mathbf{x}_c). \tag{6}$$

Finally, we apply a transformer fusion mechanism to the channel-attended features, incorporating condition features:

$$\mathbf{f}_t = \text{Transformer}(\mathbf{x}_{ca}, \mathbf{z}_t). \tag{7}$$

Specifically, $\mathbf{f}_t$ consists of a self-attention and a cross-attention, the complete $\mathbf{f}_t$ can be formulated as:

$$\mathbf{f}_t = \text{CA}(\text{SA}(\mathbf{x}_{ca}), \mathbf{z}_t). \tag{8}$$

Source	Method	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogleNet	VGG-16	ViT-B	ViF-S	Swin-T
Inc-v3	Logit	**99.60***	5.80	6.50	1.70	3.00	0.80	1.50	0.59	3.09	1.96
	SU	99.59*	5.80	7.00	3.35	3.50	2.00	3.94	0.40	1.01	0.36
	Everywhere	99.31*	50.72	45.46	34.40	42.34	40.28	41.18	5.73	27.19	17.73
	C-GSP	93.40*	66.90	**66.60**	41.60	46.40	40.00	45.00	21.18	40.86	21.48
	CGNC	96.03*	59.41	47.98	42.50	62.91	51.36	52.63	24.81	52.43	28.16
	TGAF	98.15*	**72.49**	63.20	**61.94**	**78.30**	**66.86**	**70.64**	**33.03**	**63.27**	**42.61**
Res-152	Logit	10.10	10.70	12.80	95.70*	12.70	3.70	9.20	2.09	21.25	5.19
	SU	12.36	11.31	16.16	95.08*	16.13	6.55	14.28	2.29	19.94	4.70
	Everywhere	60.38	55.20	43.00	97.66*	74.16	34.80	60.90	11.77	57.08	38.83
	C-GSP	37.70	47.60	**45.10**	93.20*	64.20	41.70	45.90	22.01	36.66	21.65
	CGNC	53.39	51.59	34.18	95.85*	85.60	62.27	63.36	34.81	58.68	40.84
	TGAF	**62.60**	**62.44**	44.02	**97.79***	**87.90**	**63.54**	**65.20**	**39.64**	**63.36**	**42.84**

Table 1: Attack success rates (%) for multi-target attacks against **normally trained** models. * represents white-box attacks.

In the equations above, **SA** (self-attention) and **CA** (cross-attention) are defined as:

Self-Attention (SA):

$$\mathrm{SA}(\mathbf{x}_{ca}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{ca}\mathbf{K}_{ca}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{ca}. \tag{9}$$

Cross-Attention (CA):

$$\mathrm{CA}(\mathbf{x}_{sa}, \mathbf{z}_t) = \mathrm{Softmax}\left(\frac{\mathbf{Q}_{sa}\mathbf{K}_{z_t}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{z_t}. \tag{10}$$

Here, $\mathbf{Q}_{ca}, \mathbf{K}_{ca}, \mathbf{V}_{ca}$ are query, key, and value matrices from the input $\mathbf{x}_{ca}$, while $\mathbf{K}_{z_t}$ and $\mathbf{V}_{z_t}$ come from the target-conditioned representation $\mathbf{z}_t$.

Image Decoder Module. The decoder $\mathcal{D}$ upsamples features from the Feature Integration Module through transposed convolutions to reconstruct $\boldsymbol{o}$. Following C-GSP (Yang et al. 2022), we apply $\tanh(\cdot)$ projection with budget ϵ to obtain the final output $\boldsymbol{\delta} = \epsilon \cdot \tanh(\boldsymbol{o})$.

Mask Mechanism. We introduce a dynamic block-wise masking strategy that randomly masks regions in the noise pattern δ. Specifically, we partition the image into $N \times N$ blocks of varying sizes and randomly select two blocks for masking.

Training Objectives. The proposed TGAF framework is trained to minimize the cross-entropy loss between the victim model's prediction and the target class label c_t through an end-to-end optimization process:

$$\theta^* \leftarrow \arg\min_{\theta} \mathcal{L}_{\mathrm{CE}}\left(f_{\Phi}(\boldsymbol{x}_s + \mathcal{D}_{\theta}(\mathcal{F}_{\theta}(\mathcal{E}_{\theta}, \mathcal{G}_{\theta}), c_t)\right). \tag{11}$$

where the adversarial sample $\boldsymbol{x}_s + \mathcal{D}_{\theta}(\cdot)$ is fed into the victim model f_{Φ} to compute the cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ that drives the attack towards the target class c_t.

4 Experiments

4.1 Experimental Settings

Datasets. Following prior works (Feng et al. 2023), the generator is trained on the ImageNet training set (Deng et al. 2009), while the performance is evaluated on the ImageNet-NeurIPS (1k) dataset (NeurIPS 2017). Additional results on ImageNet Validation Set (50k) and COCO (Lin et al. 2014) are provided in Appendices A and B, respectively.

Victim Models. We assess the effectiveness across three types of settings. 1) **Normally trained models,** including Inception-v3 (Inc-v3) (Szegedy et al. 2016), Inception-v4 (Inc-v4) (Szegedy et al. 2017), Inception-ResNet-v2 (Inc-Res-v2) (Szegedy et al. 2017), ResNet-152 (Res-152) (He et al. 2016b), DenseNet-121 (DN-121) (Huang et al. 2017), GoogleNet (Szegedy et al. 2015), VGG-16 (Simonyan and Zisserman 2014), ViT-B (Dosovitskiy et al. 2020), Visformer (ViF-S) (Chen et al. 2021) and Swin-Tiny (Swin-T) (Liu et al. 2021). 2) **Robustly trained models,** which consist of adv-Inception-v3 (Inc-v3$_{\mathrm{ADV}}$) (Goodfellow, Shlens, and Szegedy 2014), ens-adv-Inception-ResNet-v2 (IR-v2$_{\mathrm{ENS}}$) (Hang et al. 2020), and several variations of resilient ResNet-50 models (Geirhos et al. 2018; Hendrycks et al. 2019), including Res50$_{\mathrm{SIN}}$, Res50$_{\mathrm{IN}}$, Res50$_{\mathrm{FINE}}$ (Geirhos et al. 2018)), and Res50$_{\mathrm{AUG}}$ (Hendrycks et al. 2019)). **Defense methods,** consisting of three strategie. Preprocessing techniques such as JPEG compression (Dziugaite, Ghahramani, and Roy 2016), BitSqueezing (Xu, Evans, and Qi 2017), and Smoothing (Ding, Wang, and Jin 2019). Denoising methods such as the high-level representation guided denoiser (HGD) (Liao et al. 2018). Diffusion-based methods including DiffPure (Nie et al. 2022) and NRP (Naseer et al. 2020).

Baseline Methods. For instance-specific attacks, we include Logit (Zhao, Liu, and Larson 2021), SU (Wei et al. 2023), and Everywhere (Zeng et al. 2025). For SU and Everywhere, we use their best-performing variants: DTMI-Logit-SU and CFM-Everywhere, respectively. For instance-agnostic attacks, we compare against C-GSP (Yang et al. 2022) and CGNC (Fang et al. 2024). All baseline results are reproduced using the official code and weights.

Implementation Details. Following previous works (Feng et al. 2023), we use Inc-v3 and Res-152 as surrogate models for training the generator. The reported results are averaged over 8 different target classes. The perturbation budget is set to $\epsilon = 16/255$. The generator is trained for 10 epochs with a learning rate of 2e-4 and a batch size of 16. The number of blocks, N, is set to 3. Specifically, the diffusion model is used only once per target class before training begins to generate the corresponding 2D tensors, which are then saved to disk. During the training and inference of TGAF, we simply load these pre-computed tensors directly.

Source	Method	Inc-v3$_{ADV}$	IR-v2$_{ENS}$	Res50$_{SIN}$	Res50$_{IN}$	Res50$_{FINE}$	Res50$_{AUG}$
Inc-v3	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
	Everywhere	0.68	1.19	4.73	27.04	39.37	18.15
	C-GSP	20.41	18.04	6.96	33.76	44.56	21.95
	CGNC	24.30	22.51	8.88	40.81	52.13	22.83
	TGAF	**39.69**	**34.86**	**17.76**	**64.79**	**72.36**	**43.53**
Res-152	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
	Everywhere	0.55	1.23	9.71	59.94	81.45	50.09
	C-GSP	14.60	16.01	16.84	60.30	65.51	42.88
	CGNC	22.15	26.70	29.81	79.82	84.05	63.66
	TGAF	**27.73**	**32.71**	**38.07**	**84.53**	**88.48**	**68.63**

Table 2: Attack success rates (%) for multi-target attacks against **robustly trained** models.

Source	Method	Smoothing			JPEG compression			BitSqueezing			HGD
		Gaussian	Medium	Average	Q=65	Q=75	Q=85	4-Bits	5-Bits	6-Bits	
Inc-v3	CGNC	34.10	30.21	26.89	28.40	34.09	40.80	36.06	50.10	52.29	52.91
	TGAF	**36.59**	**45.23**	**36.25**	**37.93**	**46.01**	**55.90**	**51.90**	**68.36**	**70.34**	**70.49**
Res-152	CGNC	57.15	56.15	50.06	50.24	55.27	59.15	42.89	60.69	62.98	78.91
	TGAF	**58.54**	**59.08**	**51.68**	**53.46**	**57.85**	**61.22**	**48.61**	**61.90**	**64.50**	**83.02**

Table 3: Attack success rates (%) across **preprocessing technique** under VGG-16 and **HGD**.

Defense	Method	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogLeNet	VGG-16	ViT-B	ViF-S	Swin-T
Diffpure	CGNC	0.43	**0.33**	0.21	0.23	0.40	0.37	0.21	0.23	0.24	0.33
	TGAF	**0.94**	0.32	**0.28**	**0.39**	**0.46**	**0.69**	**0.61**	**0.33**	**0.51**	**0.58**
NRP	CGNC	12.73	4.25	2.28	3.10	5.16	4.51	4.01	1.61	4.33	7.43
	TGAF	**21.86**	**7.04**	**4.00**	**7.12**	**9.29**	**7.33**	**8.26**	**3.50**	**9.14**	**13.56**

Table 4: Attack success rates (%) across **Diffpure and NRP**. The surrogate model is Inc-v3.

4.2 Evaluation on Normal Models

We first assess the transferability of adversarial examples generated by normally trained models. Tab. 1 demonstrates that TGAF consistently achieves state-of-the-art (SOTA) performance, exhibiting the highest black-box ASR across nearly all experimental configurations. For examples, on CNN-based models, TGAF outperforms the previous SOTA method CGNC by 16.11% for Inc-v3 and 5.89% for Res-152, respectively. For Transformer-based models, the improvements are 11.17% for Inc-v3 and 3.84% for Res-152.

4.3 Evaluation on Robust Models

Building upon the findings obtained with normally trained models, we proceed to evaluate the performance of our method on robustly trained models. Tab. 2 shows that our method also achieves SOTA performance. Specifically, when using Inc-v3 and Res-152 as surrogate models, our approach delivers average ASR improvements of 16.92% and 5.66% over CGNC, and exceeds other methods by more than 30% and 10%, respectively.

4.4 Evaluation on Defense Methods

To further evaluate the robustness, we evaluate TGAF against multiple defense mechanisms. Tabs. 3 and 4 show that TGAF outperforms CGNC across nearly all tested defense mechanisms. It's important to note that both DiffPure and NRP significantly reduce the targeted attack success rate. This is an expected outcome, as these defenses are designed to substantially alter input information (or its internal representations) to "cleanse" adversarial perturbations. This purification process, however, is inherently lossy. While it effectively removes the perturbation, it can inadvertently corrupt the original semantic features crucial for correct classification. Therefore, a critical point must be emphasized: a low ASR does not necessarily mean the model recovered the correct classification. In many cases, the purified image, while not misclassified as the attacker's intended target, is instead misclassified as a different, incorrect class due to this information loss. Despite these robust defenses, TGAF consistently and substantially outperforms the CGNC baseline. For example, when tested against NRP on Swin-T, TGAF achieves a ASR of 13.56%, which is nearly double that of CGNC's 7.43%.

4.5 Evaluation on Adversarial Sample Quality

Tab. 8 shows that TGAF exhibits only marginal differences in SSIM, LPIPS, and FID metrics compared to CGNC. For instance, the LPIPS differs by merely 0.013. Simultane-

Source	Method	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogLeNet	VGG-16	ViT-B	ViF-S	Swin-T
Inc-v3	TGAF-C	96.23*	64.21	51.60	50.11	70.18	61.36	60.55	27.80	50.96	25.17
	TGAF	**98.15***	**72.49**	**63.20**	**61.94**	**78.30**	**66.86**	**70.64**	**33.03**	**63.27**	**42.61**
Res-152	TGAF-N	56.28	54.95	40.60	95.26*	83.28	62.95	61.84	33.53	53.28	31.76
	TGAF	**62.60**	**62.44**	**44.03**	**97.79***	**87.90**	**63.54**	**65.20**	**39.64**	**63.36**	**42.84**

Table 5: **Ablation study on masking strategies.** TGAF-C represents the substitution of our masking strategy with the CGNC masking strategy, while TGAF-N represents the removal of our masking strategy.

Method	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogleNet	VGG-16	ViT-B	ViF-S	Swin-T
TGAF	**62.60**	**62.44**	**44.02**	**97.79***	**87.90**	**63.54**	**65.20**	39.64	63.36	42.84
TGAF-Conv	59.65	60.24	38.80	97.49*	87.76	59.38	63.78	**40.80**	**65.99**	**43.09**
TGAF-CA	47.34	46.75	32.18	93.10*	76.96	53.15	56.56	28.95	45.19	27.10

Table 6: **Ablation study on fusion strategy.** TGAF-Conv represents the TGAF model without the CbF module. TGAF-CA represents the TGAF model without the TbF module. The surrogate model is Res-152.

Block (N)	Inc-v3	Inc-v4	Inc-Res-v2	Res-152	DN-121	GoogleNet	VGG-16	ViT-B	ViF-S	Swin-T
2	55.70	52.64	35.66	97.25*	84.34	56.25	57.01	32.96	58.36	40.96
3	**62.60**	**62.44**	**44.02**	**97.79***	**87.90**	63.54	65.20	**39.64**	**63.36**	**42.84**
4	60.31	59.79	40.73	97.68*	86.91	**64.95**	**68.09**	35.71	60.20	41.95

Table 7: **Analysis of block number sensitivity.** The surrogate model is Res-152.

ously, TGAF achieves a slightly better PSNR score. We contend that this negligible difference in perceptual quality represents a highly successful trade-off for a substantial improvement in attack transferability.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)
CGNC	27.986	**0.814**	**0.186**	**244.633**
TGAF	**27.994**	0.801	0.199	259.945

Table 8: **Adversarial sample quality.** (Res-152 surrogate)

4.6 Ablation Study

To thoroughly evaluate the effectiveness of TGAF, we conduct comprehensive ablation studies on masking strategies, fusion mechanisms, and block number sensitivity. Tab. 5 analyzes the effects of the masking strategy. Results demonstrate that when Inc-v3 serves as the surrogate model, our method achieves an average ASR improvement of 9.24% for CNN-based models and 11.66% for Transformer-based models. Furthermore, compared to TGAF-N with Res-152 as the surrogate model, our masking strategy leads to a enhancement of 4.31% for CNN-based models and 9.10% for Transformer-based models. Tab. 6 details the performance on fusion mechanisms. It shows that removing the TbF led to a significant performance degradation across all black-box models. For the CbF, removing it resulted in a slight decrease in ASR against most CNN models (an average drop of about 2.68%). This indicates that the CbF module positively contributes to attacking CNN models by learning local feature interactions. Interestingly, we observed that the performance of TGAF-Conv slightly improved when attacking Transformer architectures. We hypothesize that this is because CbF learns features that might slightly overfit to CNN-style architectures, thereby hindering transferability to Transformer models. Tab. 7 investigates the sensitivity to the number of blocks (N). The results show that setting $N = 3$ consistently outperforms $N = 2$ across all 10 target models. When comparing $N = 3$ and $N = 4$, we observe that while $N = 4$ performs marginally better on certain classic CNNs, $N = 3$ exhibits stronger overall performance on the majority of models. Crucially, the attack performance remains at a very high level for both $N = 3$ and $N = 4$, without any sharp drop in ASR. This robust performance indicates that the success of our method stems from its core design principles and a well-balanced architecture, rather than a fragile dependency on a highly fine-tuned hyperparameter.

5 Conclusion

In this paper, we identify that the quality and quantity of the implanted target semantic features are the key factors influencing the transferability of adversarial attacks. To address these challenges, we propose TGAF, a novel method that leverages the abilities of diffusion models to encode target labels as 2D tensors, improving the quality of target semantic information. Furthermore, we introduce a masking strategy during the training phase to ensure that the portion of the generated noise retains complete target semantic features, thus enhancing the quantity. Our extensive experiments demonstrate that TGAF not only surpasses existing SOTA methods on normally trained models but also shows robust performance against a variety of defense strategies. We hope that our proposed method serves as a reliable tool for evaluating model robustness under black-box attack settings and provides new insight for further research on vulnerability and robustness in adversarial scenarios.

Acknowledgments

This work was supported by the National Science and Technology Innovation 2030 - Major Project (Grant No. 2022ZD0208800), and NSFC General Program (Grant No. 62176215).

References

Byun, J.; Cho, S.; Kwon, M.-J.; Kim, H.-S.; and Kim, C. 2022. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15244–15253.

Byun, J.; Kwon, M.-J.; Cho, S.; Kim, Y.; and Kim, C. 2023. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24648–24657.

Chen, Z.; Xie, L.; Niu, J.; Liu, X.; Wei, L.; and Tian, Q. 2021. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 589–598.

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

Ding, G. W.; Wang, L.; and Jin, X. 2019. AdverTorch v0. 1: An adversarial robustness toolbox based on pytorch. *arXiv preprint arXiv:1902.07623*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4312–4321.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dziugaite, G. K.; Ghahramani, Z.; and Roy, D. M. 2016. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*.

Fang, H.; Kong, J.; Chen, B.; Dai, T.; Wu, H.; and Xia, S.-T. 2024. Clip-guided generative networks for transferable targeted adversarial attacks. In *European Conference on Computer Vision*, 1–19. Springer.

Feng, W.; Xu, N.; Zhang, T.; and Zhang, Y. 2023. Dynamic generative targeted attacks with pattern injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16404–16414.

Gao, L.; Cheng, Y.; Zhang, Q.; Xu, X.; and Song, J. 2021. Feature space targeted attacks by statistic alignment. *arXiv preprint arXiv:2105.11645*.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hang, J.; Han, K.; Chen, H.; and Li, Y. 2020. Ensemble adversarial black-box attacks against deep learning systems. *Pattern Recognition*, 101: 107184.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. 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*, 630–645. Springer.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Kong, Z.; Guo, J.; Li, A.; and Liu, C. 2020. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14254–14263.

Li, M.; Yang, Y.; Wei, K.; Yang, X.; and Huang, H. 2022. Learning universal adversarial perturbation by adversarial example. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 1350–1358.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1778–1787.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, H.; Peng, B.; Ding, P.; and Wang, D. 2025. Enhancing Adversarial Transferability via Component-Wise Augmentation Method. *arXiv preprint arXiv:2501.11901*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of*

the IEEE/CVF international conference on computer vision, 10012–10022.

Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; and Lu, F. 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110: 107332.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2020. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 262–271.

Naseer, M.; Khan, S.; Hayat, M.; Khan, F. S.; and Porikli, F. 2021. On generating transferable targeted perturbations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7708–7717.

Naseer, M. M.; Khan, S. H.; Khan, M. H.; Shahbaz Khan, F.; and Porikli, F. 2019. Cross-domain transferability of adversarial perturbations. *Advances in Neural Information Processing Systems*, 32.

NeurIPS. 2017. https://www.kaggle.com/c/nips-2017-defense-against-adversarial-attack/data.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Wang, K.; He, X.; Wang, W.; and Wang, X. 2024. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24336–24346.

Wang, X.; and He, K. 2021. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1924–1933.

Wei, Z.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2023. Enhancing the self-universality for transferable targeted attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12281–12290.

Weng, J.; Luo, Z.; Li, S.; Sebe, N.; and Zhong, Z. 2023. Logit margin matters: Improving transferable targeted adversarial attack by logit calibration. *IEEE Transactions on Information Forensics and Security*, 18: 3561–3574.

Wu, H.; Ou, G.; Wu, W.; and Zheng, Z. 2024. Improving transferable targeted adversarial attacks with model self-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24615–24624.

Xu, W.; Evans, D.; and Qi, Y. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*.

Yang, X.; Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2022. Boosting transferability of targeted adversarial examples via hierarchical generative networks. In *European conference on computer vision*, 725–742. Springer.

Zeng, H.; Cui, S.; Chen, B.; and Peng, A. 2025. Everywhere Attack: Attacking Locally and Globally to Boost Targeted Transferability. *arXiv preprint arXiv:2501.00707*.

Zhao, A.; Chu, T.; Liu, Y.; Li, W.; Li, J.; and Duan, L. 2023. Minimizing maximum model discrepancy for transferable black-box targeted attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8153–8162.

Zhao, Z.; Liu, Z.; and Larson, M. 2021. On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems*, 34: 6115–6128.