Boundary Box-Guided Targeted Adversarial Attacks with Semantic Perturbation

Hongtian Zhao[1] and Wenzhuo Shi[2] and Chang Liu[3] and Yiquan Wang[4]

Abstract— **Targeted adversarial attacks in black-box settings are pivotal for uncovering vulnerabilities in neural networks and guiding the development of robust defenses. However, conventional attack methods typically perturb the primary content, leading to a degradation in image quality and highlighting the need for more reasonable optimization strategies. In contrast, we propose a novel algorithm that restricts perturbations to the image boundary regions, thereby preserving content fidelity while enhancing attack effectiveness. Our approach employs an encoder–decoder generative network to craft targeted adversarial examples guided by optimized semantic perturbations derived from the boundaries. Moreover, the boundary signal is jointly optimized with the model parameters, enabling efficient, amortized optimization for multi-class targeted attacks. Extensive experiments demonstrate that the proposed boundary-guided method significantly improves the success rates of targeted black-box attacks and can be seamlessly integrated into existing noise-injection techniques to enhance overall performance.**

I. Introduction

In deep learning-based intelligent systems, adversarial attacks have become a central concern as they exploit subtle, often imperceptible perturbations that dramatically alter neural network predictions [1], [2]. Targeted attacks, which mislead models into predicting specific adversary-chosen classes, have gained significant attention due to their practical implications and efficiency compared to untargeted approaches [3]. These vulnerabilities raise serious concerns for AI deployment in critical domains including autonomous vehicles [4], healthcare [5], and financial systems [6], where successful attacks could have severe consequences. Black-box attacks, operating without access to model parameters, exploit the transferability of adversarial examples across different models. This transferability phenomenon not only reveals shared feature representations and decision boundaries but also informs the development of robust defense strategies.

In recent years, targeted adversarial attack techniques have evolved into two main categories: optimization-based and generation-based approaches. Optimization-based methods [2], [7]–[9] directly optimize perturbations using surrogate models through iterative gradient-based techniques, but often suffer from poor transferability in black-box scenarios due to overfitting [2], [10]. Generation-based methods [3], [11] employ dedicated models that learn universal mapping functions to produce adversarial perturbations from normal inputs. By training on unlabeled datasets, these approaches generate more transferable examples by reducing data-specific overfitting. CGSP [11], a representative generation-based method, enhances transferability through hierarchical networks that minimize dependency on individual samples. Despite these advances, both approaches share a critical limitation: they introduce perturbations directly into primary image content, degrading visual quality and compromising attack stealthiness. This limitation is particularly problematic in real-world scenarios where adversarial examples must remain imperceptible to avoid detection or preserve content integrity, such as in security systems where attackers might inject adversarial noise along image boundaries to evade detection or mislead classifiers.

To address these limitations, we depart from conventional methods that focus on perturbing image content. Instead, we propose a novel approach that enhances adversarial attack efficacy through the integration of external encrypted signals. This strategy not only improves attack performance but also challenges emerging defense mechanisms, necessitating the development of more sophisticated defensive countermeasures. Our investigation is guided by the following research questions:"*Can adversarial attack performance be improved by leveraging external signals rather than relying solely on intrinsic image noise? Furthermore, can the integration of external signals with existing targeted attack methods enhance overall attack efficacy?*"

To address these challenges, we propose a boundary-guided perturbation framework that shifts focus from altering primary image content to optimizing noise in independent, non-overlapping external regions. These peripheral areas (e.g., outer edges or margins) lack semantic content, allowing perturbations to be applied without affecting human perception or interpretation of the original image. Experimental results demonstrate that injecting optimized external encrypted signals into these boundary regions significantly enhances targeted adversarial attack effectiveness. Furthermore, because boundary-based noise operates as an independent signal distinct from content perturbations, it can be seamlessly integrated with existing targeted attack methods to boost their performance.

We propose a streamlined generative framework for creating targeted adversarial examples. This framework incorporates an encoder-decoder architecture with a specialized mapping structure to optimize boundary signals. By training on an unlabeled clean image set that mirrors the original data distribution, our method reduces dependency on individual samples, enabling the model to learn semantic patterns

[1]Hongtian Zhao, [2]Wenzhuo Shi and [4]Yiquan Wang are with the Xinjiang University, Urumqi, China zhaohongtian@xju.edu.cn

[3]Chang Liu is with the Tsinghua University, Beijing, China liuchang6513@mail.tsinghua.edu.cn

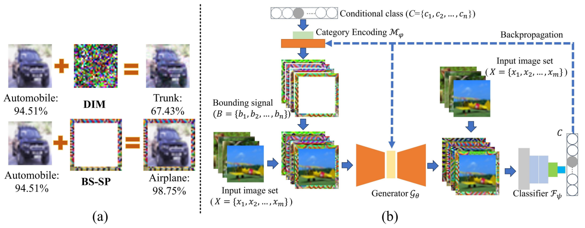
Automobile:
94.51%
DIM
Trunk:
67.43%
Automobile:
94.51%
BS-SP
Airplane:
98.75%
(a)
Conditional class (C={$c_1, c_2, \ldots, c_n$})
Category Encoding $\mathcal{M}_\varphi$
Backpropagation
Input image set
($X = \{x_1, x_2, \ldots, x_m\}$)
Bounding signal
($B = \{b_1, b_2, \ldots, b_n\}$)
Input image set
($X = \{x_1, x_2, \ldots, x_m\}$)
Generator $\mathcal{G}_\theta$
C
Classifier $\mathcal{F}_\psi$
(b)

Fig. 1: (a) shows targeted adversarial examples created using DIM [10] and BS-SP with a maximum perturbation of $\epsilon = 16$ and bounding box parameter $h = 3$. Predicted labels and probabilities are from a black-box model. (b) illustrates our generative approach for BS-SP, comprising a conditional generator and classifier. The generator integrates the input image with a conditional class vector from the category encoding structure. During training, only the generator and category encoder are optimized to probe the target classifier's decision boundaries.

more effectively. This approach significantly enhances the transferability and success rate of targeted black-box attacks. Our boundary signal optimization scheme is fully compatible with existing content attack methods, thereby extending its applicability across diverse scenarios.

In summary, this work presents three primary contributions.

- We propose a novel targeted adversarial attack algorithm that generates semantically coherent perturbations by leveraging image boundary regions and an unlabeled clean image set, which approximates the training data distribution to enhance the attack's effectiveness and generalization;
- We develop a transferable and efficient framework that integrates an encoder-decoder generative network and a mapping structure to jointly optimize boundary signals, enabling amortized optimization for multi-class targeted attacks without parameter tuning;
- The proposed method exhibits robust scalability, as it can be readily integrated with conventional adversarial perturbation generation techniques. Experimental results validate the superiority of our approach over existing methods in both effectiveness and generability.

II. Method

The proposed framework improves attack effectiveness and transferability by generating and injecting an external encrypted perturbation signal defined over bounding boxes rather than directly altering image content. In this section, we formally define the external perturbation signal and present a conditional generative model that learns a universal adversarial function over bounding boxes, enabling effective multi-target attacks in black-box settings.

A. Problem Formulation

We define the external signal as a perturbation, termed Bounded Box Signal for Semantic Perturbation (BS-SP), which is injected outside the image content as illustrated in Fig. 1a. Fig. 1b outlines the optimization process for generating a bounding box signal for multi-target attacks. In the signal injection module, a bounding box signal of width h is applied along the image boundary. This strategic placement preserves the original image content while enabling adversarial capabilities. Direct injection into the image would compromise content integrity and potentially reduce attack effectiveness.

In this study, x_s denotes an image drawn from an unlabeled dataset $\mathcal{X}_s$, a subset of $\mathbb{R}^d$, while c represents a selected target category from the set $\mathcal{C}$. Let $\mathcal{F}_\psi : \mathcal{X}_s \to \mathbb{R}^K$ be a classifier that outputs a probability vector corresponding to K categories. The generation of a targeted adversarial example, x_s^*, from an authentic sample x_s entails manipulating the classification model $\mathcal{F}_\psi$ such that it classifies x_s^* as category c. Here, c is determined by the expression $\arg\max_{i \in \mathcal{C}} \mathcal{F}_\psi(x_s^*)_i = c$. The adversarial example x_s^* can be expressed as $\mathcal{B}(x_s, b)$, where $\mathcal{B}(\cdot)$ incorporates the bounding box signal b into the image, with b representing the encrypted signal associated with x_s.

While some generative approaches [12], [13] are capable of learning targeted adversarial perturbations, they overlook the efficiency of generating perturbations for multiple targets, resulting in practical limitations. To address this limitation, we propose a conditional generative network $\mathcal{G}_\theta$ that is specifically designed to efficiently generate adversarial perturbations targeting multiple classes through class-conditional distributions. In contrast to previous single-target methods [12], [13], the proposed framework treats the target label c as a variable rather than a fixed constant. As illustrated in Fig. 1b, our framework consists of a conditional generator $\mathcal{G}_\theta$, a category encoder $\mathcal{M}_\varphi$, and a classifier network $\mathcal{F}_\psi$, each parameterized by θ, φ, and ψ, respectively. The category encoder $\mathcal{M}_\varphi$ converts the category label c into a bounding box tensor denoted as b_c. In this framework, the conditional generative model $\mathcal{G}_\theta : (\mathcal{X}_s, \mathcal{B}) \to P$ generates a perturbation $\delta = \mathcal{G}_\theta(x_s, b_c)$ that resides within the space $P \subset \mathbb{R}^e$, as defined by the training dataset. The output δ from $\mathcal{G}_\theta$ is constrained to a specified border width h, thereby producing the altered image defined as $x_s^* = \mathcal{B}(x_s, \delta)$.

Utilizing a pre-trained network $\mathcal{F}_\psi$, we aim to produce targeted adversarial perturbations by addressing the following optimization problem:

$$\arg\min_{\theta,\varphi} \mathbb{E}_{(x_s \sim \mathcal{X}_s, c \sim \mathcal{C})} [\mathbb{CE}(\mathcal{F}_\psi(\mathcal{G}_\theta(\mathcal{B}(x_s, \mathcal{M}_\varphi(c)))), c)],$$

.

$$s.t. \quad W((\mathcal{M}_\varphi(c))) = h. \tag{1}$$

Here, $\mathbb{CE}$ represents the cross-entropy loss. $W((\mathcal{M}_\varphi(c)))$ is the width of the border. Addressing this challenge enables the development of a targeted conditional generator that minimizes the loss for a specific target class within the unlabeled training dataset. It is noteworthy that we refine the parameters θ, φ of the generator $\mathcal{G}_\theta$, $\mathcal{M}_\varphi$ with the training dataset $\mathcal{X}_c$. Subsequently, for any specific image x_t in the test dataset $\mathcal{X}_t$, the targeted adversarial perturbation δ_t can be formulated as $\delta_t = \mathcal{G}_\theta(\mathcal{B}(x_t, \mathcal{M}_\varphi(c)))$, $x_t^* = \mathcal{B}(x_t, \delta_t)$. This formulation necessitates only a single inference for the targeted image x_t.

The experimental analysis reveals that the objective function in Equation 1 significantly enhances the transferability of the crafted perturbation δ. This improvement likely stems from δ capturing semantic features inherent to the target class, making it robust against variations in the training data. Notably, our findings demonstrate that the semantic patterns produced by the proposed method are consistently classified with high confidence as the intended target class. In contrast, perturbations generated by DIM [10] predominantly resemble random noise without semantic coherence.

B. Network Architecture and Algorithm

We present a conditional adversarial generative framework specifically designed for targeted attacks, as illustrated in Fig. 1b. In particular, we construct a mapping network $\mathcal{M}_\varphi$ that produces a vector specific to each target within the implicit space, and we train the conditional generator $\mathcal{G}_\theta$ to emulate this vector, perpetually deceiving the classifier $\mathcal{F}_\psi$.

Mapping network. With the one-hot class encoding $\mathbf{l}_c \in \mathbb{R}^K$ corresponding to a specific target class c, the mapping network is designed to produce a specific latent vector $w = \mathcal{M}_\varphi(\mathbf{l}_c)$, where $w \in \mathbb{R}^M$. The function $\mathcal{M}_\varphi(\cdot)$ includes a multi-layer perceptron (MLP) and a normalization layer, enabling the creation of varied targeted vectors w for any designated target class c. Consequently, $\mathcal{M}_\varphi$ has the ability to acquire efficient targeted latent vectors through the random selection of various classes $c \in C$ during the training period.

Generator. For the VAE generator, we introduce a hierarchical learning architecture that leverages a VQ-VAE-based backbone [14]. Unlike existing methods, following the downsampling operation, the latent representation is processed through two dense blocks [15]. Each dense block comprises five layers, with each layer employing 256 filters and a growth rate of 32, serving as interim fusion modules to enhance feature integration within the latent space. Formally, upon receiving an input image x_s, the encoder initially computes the feature map $F \in R^{N\times H\times L}$, with N, H, and L indicating the number of channels, the feature map's height, and its width, respectively. The target latent vector w, produced by the mapping network $\mathcal{M}_\varphi$ through the introduction of a designated target class c, is extended across both the height and width to create the label feature map $w_s \in R^{M\times H\times L}$. Subsequently, these two feature maps are merged across the channel dimension to form $F^{'} \in R^{(N+M)\times H\times L}$. This combined feature map is then supplied to the next network layer. Hence, our generator $\mathcal{G}_\theta$ transforms an input image x_s and a latent target vector w into an output image $\mathcal{G}_\theta(x_s, w)$. This transformation allows $\mathcal{G}_\theta$ to generate adversarial images targeting a range of classes. Given the border width h, the final dimension of feature vector can be represented as $F \in R^{M\times h\times(2\times(H+L)+4\times h)}$. To optimize the smoothness and standardization of the output, we apply the hyperbolic tangent function as a nonlinear transformation to the generated feature vectors:

$$\delta = \tanh(f), \tag{2}$$

the activation function naturally fits this scenario as it outputs values between $[-1, 1]$, ensuring the preservation of data integrity and eliminating the need for additional clipping steps. Its zero-centered output and symmetric range help stabilize training by promoting even weight updates and reducing bias, which aligns well with the standardized scale of image pixels, ensuring that the generated images are immediately usable. Training Goals. The primary aims of the training process are to reduce the classification inaccuracies associated with the images with border signal produced by the generator as

$$\theta^*, \varphi^* \leftarrow \arg\min_{\theta,\varphi} \mathbb{CE}(\mathcal{F}_\psi(\mathcal{G}_\theta(\mathcal{B}(x_s, \mathcal{M}_\varphi(c)))), c). \tag{3}$$

This approach employs a comprehensive training strategy to generate adversarial examples that mislead classifiers toward the target label, utilizing cross-entropy loss ($\mathbb{CE}$) as the optimization objective. The complete optimization procedure is detailed in Algorithm 1.

Algorithm 1 Training Algorithm for the Conditional Bounding Adversarial Generative Networks

```
Require: Training data $\mathcal{D}_s$, a generative network $\mathcal{G}_\theta$, a classifier $\mathcal{F}_\psi$, a category mapping structure $\mathcal{M}_\varphi$, maximum number of iterations $N$.
Ensure: Parameters of the adversarial signal generator for bounding boxes $\theta$, $\varphi$.
while Iter $\leq N$ do
    Randomly sample a batch of images, $x_s$, from the set $\mathcal{X}_s$;
    Randomly sample a batch of target labels, $c$, from the set $\mathcal{C}$;
    Forward pass the target labels $c$ into the mapping structure $\mathcal{M}_\varphi$ to compute the latent category vectors $b$;
    Obtain the adversarial samples by $x_s^* = \mathcal{B}(x_s, \tanh(\mathcal{G}_\theta(\mathcal{B}(x_s, b))))$;
    Forward pass the adversarial samples $x_s^*$ to the discriminator $\mathcal{F}_\psi$ and compute the loss as specified in Equation 3;
    Perform a backward pass and update the parameters of $\mathcal{G}_\theta$ and $\mathcal{M}_\varphi$;
end while
return $\theta$, $\varphi$
```

C. Boundary Box-Guided Encrypted Signals for Targeted Adversarial Attacks

We propose an unsupervised framework that generates imperceptible adversarial patterns with target-specific semantic information using only clean samples and target class labels. By strategically embedding these patterns along image boundaries, our method enables effective targeted attacks while preserving the visual integrity of the main content. During inference, these adversarial boundary patterns are incorporated into clean or adversarial samples, exploiting model vulnerabilities to peripheral perturbations. It complements existing adversarial techniques and integrates seamlessly with them, creating more potent attack vectors while revealing previously unexplored vulnerabilities in current defense mechanisms.

III. Experiments

In this section, we evaluate the proposed method's effectiveness in executing targeted adversarial attacks through comprehensive experiments. We implemented our approach using PyTorch and conducted evaluations on a server equipped with Intel(R) Xeon(R) Platinum P-8136 CPU @ 2.00GHz, Nvidia GeForce RTX 4090 D, and 512GB RAM. For model optimization, we trained our network for 200 epochs using the Adam optimizer with an initial learning rate of 0.015 for CIFAR-10 [16] and 2.8×10^{-3} for Tiny-ImageNet [17], with a consistent weight decay of 1×10^{-5} across all experiments. We employed a step decay schedule that reduced the learning rate by a factor of 0.4 every 50 epochs to enhance convergence stability. To improve adversarial transferability against defense-aware models, we incorporated an adaptive Gaussian smoothing mechanism during training when computing the perturbation δ from Eq. 2, thereby enhancing the robustness of generated adversarial examples across diverse defense scenarios.

A. Experiment Setup

We evaluate our boundary perturbation attack framework on two standard benchmarks: CIFAR-10 (32×32 pixel images across 10 classes) and Tiny-ImageNet (64×64 pixel images spanning 200 classes). These datasets balance computational tractability with real-world relevance, facilitating comprehensive evaluation across diverse model architectures. We scale the boundary perturbation width proportionally to each dataset's resolution: $h = 5$ pixels for CIFAR-10 and $h = 10$ pixels for Tiny-ImageNet, maintaining a consistent relative boundary proportion.

We evaluate our proposed method using diverse pre-trained architectures: ResNet-18/34/50 [18], VGG-16 [19], DenseNet-161/169 [15], and Inception-v3 [20], establishing a comprehensive baseline. For the CIFAR-10 experiments, we leverage publicly available pre-trained models[1], which are well-suited as guides for training generative models due to their proven generalizability and high performance. For Tiny-ImageNet experiments, we fine-tune standard PyTorch models to optimize classification performance specifically for this dataset, ensuring robust feature extraction during adversarial training.We benchmark our BS-SP method against established sample-specific adversarial attacks, including MIM [2], DIM [10], TIM [21], EOTPGD [22], FAB [23], NIFGSM [24], and Jitter [25]. We also compare against CGSP [11], a state-of-the-art category-specific targeted attack, to evaluate our method's robustness. All competing sample-specific attacks are implemented with optimal hyperparameters as recommended in the Torchattacks repository [26], a standard PyTorch framework for adversarial example generation.

We configure the attack parameters as follows: MIM, TIM, DIM, FAB, and Jitter use 10 iterations, EOTPGD employs 20 iterations, and NIFGSM runs for 300 iterations. All adversarial models are implemented using the Torchattacks library [26].

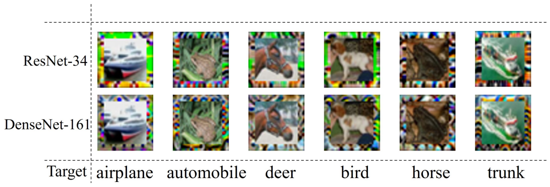

Fig. 2: The figure displays adversarial examples generated with perturbation budget $h = 5$. We use ResNet-34 and DenseNet-161 as guiding discriminator models in separate experiments to generate the targeted perturbations.

B. Experiments on CIFAR-10

To evaluate our multi-target black-box attack protocol, we conduct experiments on CIFAR-10 using six representative classes: airplane, automobile, bird, deer, horse, and truck. All generation tasks are performed on the standard 10,000-image test set following established evaluation protocols.

We analyze the computational complexity of various attack methods. Sample-specific attacks (MIM, DIM, TIM, EOTPGD, FAB, NIFGSM, and Jitter) require iterative optimization with K gradient computation steps. Their per-instance computational cost is $T_{SS} = (t_{FP} + t_{BP}) \times K$, where t_{FP} and t_{BP} represent forward and backward propagation times through the target classifier, as shown in Table I. The total computational burden varies further based on each method's search strategy. In contrast, sample-agnostic approaches require only a single forward pass through the pre-trained generator, with complexity $T_{SA} = t_{GFP}$. Table I demonstrates that both sample-agnostic methods achieve comparable inference efficiency while providing substantial speedup over sample-specific ones, making them suitable for time-sensitive applications.

Table I compares the effectiveness of various methods against naturally pre-trained models. Sample-specific attacks show low success rates, likely due to overfitting to individual samples, which limits their cross-model generalizability. While the sample-agnostic CGSP attack achieves acceptable performance, our proposed method demonstrates superior black-box effectiveness. This performance difference stems from distinct optimization strategies: CGSP applies global clipping projections across the entire image space, whereas our approach focuses on localized perturbations within constrained bounding box regions. Our framework offers a novel perspective for targeted black-box attacks, delivering competitive performance in both attack success rate and computational efficiency.

Besides, Fig. 2 displays several targeted adversarial samples generated by the proposed method, where the boundary patterns are specifically designed to evoke semantic patterns associated with the target class.

C. Experiments on Tiny-ImageNet

To assess our method's generalization capability, we extend evaluation to the Tiny-ImageNet dataset, conducting

[1] https://github.com/huyvnphan/PyTorch_CIFAR10

TABLE I: Transferability of multi-target adversarial attacks on CIFAR-10 test set with perturbation constraint $\ell_\infty \leq 16$. Results are averaged across six target classes. Unlike sample-specific methods requiring iterative optimization for each example, our approach employs a single conditional generative model that generalizes across multiple targets, enabling efficient adversarial example generation.

	Method	Time	ResNet-18	ResNet-34	ResNet-50	VGG-16	DenseNet-161	Inception-v3
ResNet-34	MIM [2]	24.60	24.26	33.93	23.38	23.83	24.34	23.19
	TIM [21]	26.34	12.68	14.49	12.77	12.27	12.72	12.37
	DIM [10]	26.13	19.50	25.84	18.79	18.53	18.80	18.54
	FAB [23]	48.52	10.01	9.96	9.99	10.03	9.99	10.03
	Jitter [25]	28.50	10.79	11.65	11.00	10.86	10.95	10.81
	EOTPGD [22]	101.11	27.37	41.49	26.64	27.09	26.91	26.53
	NIFGSM [24]	737.04	18.46	30.21	17.95	18.08	17.30	18.18
	CGSP [11]	**0.76**	35.12	56.93	41.42	35.76	35.66	33.83
	Ours	1.06	**85.10**	**100.00**	**76.88**	**45.21**	**37.66**	**34.20**
DenseNet-161	MIM [2]	81.46	20.40	21.23	21.33	20.95	23.66	22.21
	TIM [21]	112.84	12.23	12.82	12.44	11.85	13.54	12.26
	DIM [10]	114.85	18.50	18.59	18.73	18.22	24.31	18.66
	FAB [23]	209.22	10.01	9.96	9.99	10.03	9.99	10.03
	Jitter [25]	114.78	11.51	11.50	11.74	11.76	12.96	12.15
	EOTPGD [22]	441.53	23.23	23.17	23.66	24.23	33.37	24.96
	NIFGSM [24]	3323.50	14.95	15.51	16.67	13.96	26.28	14.38
	CGSP [11]	**0.77**	30.77	36.15	41.14	28.92	57.73	31.86
	Ours	1.09	**78.45**	**85.76**	**94.59**	**56.07**	**99.70**	**62.39**

TABLE II: Targeted adversarial attack success rates on clean Tiny-ImageNet samples using DenseNet-169 and ResNet-34 as source models. Prefixes V-, D-, and R- denote VGG, DenseNet, and ResNet architectures, respectively.

	V-16	R-18	R-34	R-50	D-169
R-34	4.80	17.63	85.68	23.28	3.62
D-169	14.85	6.63	11.93	13.79	98.70

multi-target black-box attacks across eight diverse categories: Egyptian cat, beer bottle, slug, ox, fountain, flagpole, centipede, and bucket. Performance is measured by the average attack success rate across these categories.

*1) **Experiments on Normal Samples:*** We first evaluate our method on standard Tiny-ImageNet samples, using DenseNet-169 and ResNet-34 as training targets for our generative network. The generated adversarial samples are tested against multiple architectures (VGG-16, ResNet-18, ResNet-34, ResNet-50, and DenseNet-169) to assess transferability. As shown in Table II, our method achieves strong performance on clean samples and high attack success rates against surrogate networks, though cross-architecture transferability remains limited. We address this limitation through our ensemble-based approach, which significantly enhances attack generalization across diverse model architectures.

*2) **Experiments on Sample-Specific Targeted Attacks:*** As our boundary box-guided encrypted signals are sample-agnostic, they can enhance targeted adversarial attacks on any input image once optimized. To validate this generalizability, we tested our method on Tiny-ImageNet using adversarial examples generated by MIM [2] and NIFGSM [24]. We evaluated the enhanced attacks across VGG-16, ResNet-34, and DenseNet-169 models. Table III presents the results, while Fig. 3 uses heatmaps to visualize how encrypted boundaries affect model activation features, revealing their internal response mechanisms. The original adversarial samples showed limited transferability, performing well only on their source networks. However, after integrating our boundary-guided encrypted signals, attack efficacy improved significantly across models, demonstrating its effectiveness.

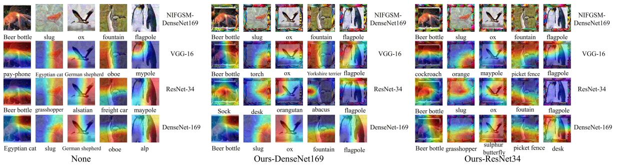

Fig. 3: The figure shows randomly selected adversarial examples generated with and without our method, along with their corresponding GradCAM heatmaps from various pre-trained models. These samples directly correspond to the NIFGSM-DenseNet169 experimental setting reported in Table III.

*3) **Experiments on Sample-Agnostic Targeted Attacks:*** We evaluate whether our method can enhance state-of-the-art sample-agnostic targeted attacks by integrating boundary box-encrypted signals. We first train our generator on clean samples, then apply it to adversarial samples generated by CGSP [11]. Table IV presents the attack success rates across methods. The results confirm that sample-agnostic targeted attacks offer significantly better generalizability than sample-specific approaches. Furthermore, incorporating our sample-agnostic boundary box-encrypted signals further enhances both attack effectiveness and transferability. This validation suggests a promising future research direction: incorporating boundary attack samples into defense model training to establish more challenging robustness benchmarks.

IV. Conclusion

In this paper, we present a novel adversarial attack framework that utilizes an external boundary box-guided encrypted signal, departing from conventional approaches that directly manipulate image content. Our method strategically embeds this external signal to precisely target specific objects while preserving the original image content. The optimized boundary signal demonstrates sample-agnostic properties, enabling universal deployment across diverse images. Furthermore, decoupling the attack signal from image content facilitates joint optimization with existing adversarial techniques. Experimental results demonstrate that our approach achieves

TABLE III: Experiments on Tiny-ImageNet samples attacked by MIM and NIFGSM methods. 'MIM-DenseNet169' indicates adversarial examples generated using MIM attack on DenseNet169, while 'Ours-DenseNet169' refers to examples produced by our approach using DenseNet169 to generate boundary box-guided signals.

	Boundary generator	VGG-16	ResNet-18	ResNet-34	ResNet-50	DenseNet-169
MIM-DenseNet169	None	3.75	5.57	5.99	6.14	62.47
	Ours-DenseNet169	**31.22**	14.78	22.37	27.84	**99.76**
	Ours-ResNet34	14.83	**29.14**	**90.50**	**41.90**	23.57
MIM-ResNet34	None	3.59	6.21	19.88	5.58	5.00
	Ours-DenseNet169	**30.34**	14.55	23.20	26.84	**99.47**
	Ours-ResNet34	14.17	**28.89**	**90.73**	**40.02**	11.39
NIFGSM-DenseNet169	None	2.41	2.98	3.24	3.21	31.34
	Ours-DenseNet169	**28.02**	12.75	20.56	25.45	**99.66**
	Ours-ResNet34	12.26	**25.62**	**89.88**	**37.35**	15.41
NIFGSM-ResNet34	None	2.81	4.41	17.42	4.07	3.72
	Ours-DenseNet169	**29.27**	14.26	22.83	26.23	**99.43**
	Ours-ResNet34	13.31	**27.51**	**90.10**	**38.95**	10.63

TABLE IV: Experiments on Tiny-ImageNet samples attacked by the CGSP method. Here, 'CGSP-DenseNet169' denotes adversarial examples generated on DenseNet169 using the CGSP attack, while 'Ours-DenseNet169' represents examples produced by our approach, which is based on DenseNet169 to generate boundary box-guided signals.

	Boundary generator	VGG-16	ResNet-18	ResNet-34	ResNet-50	DenseNet-169
CGSP-DenseNet169	None	21.50	24.49	22.53	23.14	45.63
	Ours-DenseNet169	**52.47**	27.07	30.33	41.97	**99.90**
	Ours-ResNet34	32.02	**40.14**	**92.40**	**54.23**	32.82
CGSP-ResNet34	None	20.08	26.32	33.44	20.43	18.65
	Ours-DenseNet169	**50.76**	28.98	35.49	38.79	**99.74**
	Ours-ResNet34	29.61	**42.53**	**92.86**	**53.94**	22.55
CGSP-VGG16	None	55.54	13.78	12.90	11.55	11.08
	Ours-DenseNet169	**66.21**	23.32	29.77	35.34	**99.81**
	Ours-ResNet34	47.81	**37.25**	**92.33**	**51.16**	18.47

competitive attack success rates while maintaining impressive computational efficiency compared to existing methods. Its scalability and versatility are validated through seamless integration with various content-based adversarial strategies.

Acknowledgments: This work was supported in part by the Computing and Data Center of Xinjiang University, by the National Natural Science Foundation of China (NSFC, Grant No. 12461058), by the Xinjiang Key Laboratory of Applied Mathematics (No. XJDX1401), by the Xinjiang Tianchi Talent Program "Robust Perception and Restoration of Low-Quality Visual Content" (No. 51052501848).

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. ICLR*, 2015.

[2] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *Proc. CVPR*, 2018, pp. 9185–9193.

[3] C. Zhang, P. Benz, T. Imtiaz, and I. S. Kweon, "Understanding adversarial examples from the mutual influence of images and perturbations," in *Proc. CVPR*, 2020, pp. 14 509–14 518.

[4] Z. Zhu *et al.*, "Understanding the robustness of 3d object detection with bird'view representations in autonomous driving," in *Proc. CVPR*, 2023, pp. 21 600–21 610.

[5] G. Bortsova *et al.*, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Med. Image Anal.*, vol. 73, p. 102141, 2021.

[6] M.-Y. Tsai *et al.*, " Effective Adversarial Examples Identification of Credit Card Transactions ," *IEEE Intell. Syst.*, vol. 39, no. 04, pp. 50–59, Jul. 2024.

[7] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. ICLR*, 2017.

[8] M. Li *et al.*, "Towards transferable targeted attack," in *Proc. CVPR*, 2020, pp. 638–646.

[9] J. Gu *et al.*, "A survey on transferability of adversarial examples across deep neural networks," *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

[10] C. Xie *et al.*, "Improving transferability of adversarial examples with input diversity," in *Proc. CVPR*, June 2019, pp. 2725–2734.

[11] X. Yang, Y. Dong, T. Pang, H. Su, and J. Zhu, "Boosting transferability of targeted adversarial examples via hierarchical generative networks," in *Proc. ECCV*, 2022, pp. 725–742.

[12] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proc. CVPR*, 2018, pp. 4422–4431.

[13] M. Naseer *et al.*, "Cross-domain transferability of adversarial perturbations," in *Proc. NeurIPS*, 2019, pp. 12 885–12 895.

[14] A. Razavi *et al.*, "Generating diverse high-fidelity images with VQ-VAE-2," in *Proc. NeurIPS*, 2019, pp. 14 837–14 847.

[15] G. Huang, Z. Liu, G. Pleiss, L. v. d. Maaten, and K. Q. Weinberger, "Convolutional networks with dense connectivity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8704–8716, 2022.

[16] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[17] Y. Le *et al.*, "Tiny imagenet visual recognition challenge," 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:16664790

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, June 2016, pp. 770–778.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.

[20] C. Szegedy *et al.*, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 2818–2826.

[21] Y. Dong *et al.*, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. CVPR*, 2019, pp. 4307–4316.

[22] X. Liu *et al.*, "Adv-bnn: Improved adversarial defense through robust bayesian neural network," in *Proc. ICLR*, 2019.

[23] F. Croce *et al.*, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proc. ICML*, 2020, pp. 2196–2205.

[24] J. Lin *et al.*, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. ICLR*, 2020.

[25] L. Schwinn *et al.*, "Exploring misclassifications of robust neural networks to enhance adversarial attacks," *Appl. Intell.*, vol. 53, no. 17, pp. 19 843–19 859, 2023.

[26] H. Kim, "Torchattacks: A pytorch repository for adversarial attacks," *arXiv preprint arXiv:2010.01950*, 2020.