SpAdvGAN: Learning Imperceptible Adversarial Perturbations via Sparse Semantic Attention

1st Delong Yang
Baotou Vocational & Technical College
Baotou, China
923430367@qq.com
ORCID: https://orcid.org/0009-0004-5917-9436

2nd Jiajie Xing*
Inner Mongolia University
Hohhot, China
22409026@mail.imu.edu.cn
ORCID: https://orcid.org/0009-0008-2055-2263

***Abstract*—Adversarial attacks pose serious security risks to deep neural networks (DNNs) by inducing erroneous predictions. While gradient-based and GAN-based attack methods have achieved high success rates, they often produce perturbations that are diffusely distributed across images, resulting in poor imperceptibility. This limitation restricts their applicability in scenarios such as privacy protection for face recognition systems. To address this issue, we propose SpAdvGAN, a semantic-aware adversarial example generation framework that integrates sparse attention with generative adversarial networks. SpAdvGAN introduces a sparse semantic attention extractor that explicitly constrains perturbations to a limited set of semantically important regions, preventing unnecessary noise diffusion and preserving visual fidelity. This design enables precise regulation of perturbation granularity while maintaining strong attack effectiveness. Extensive experiments on CIFAR-10 and ImageNet datasets demonstrate that SpAdvGAN substantially improves the imperceptibility of adversarial examples, while achieving comparable or higher attack success rates than state-of-the-art gradient-based attacks and the GAN-based baseline LpAdvGAN.**

***Index Terms*—Adversarial Examples; Adversarial Attacks; Generative Adversarial Networks; Sparse Attention; Imperceptible Perturbations**

I. Introduction

Deep learning models have achieved remarkable success across a wide range of computer vision tasks. However, numerous studies have demonstrated that deep neural networks (DNNs) remain highly vulnerable to adversarial attacks, where carefully crafted perturbations can induce incorrect predictions while remaining visually inconspicuous to human observers [1], [2]. Such adversarial examples pose serious security threats in safety-critical applications, including biometric authentication systems and autonomous driving scenarios [3], [4]. Consequently, understanding how to generate effective adversarial examples has become a central topic in the study of model robustness.

Existing adversarial example generation methods can be broadly categorized into three classes. Gradient-based methods, such as the Fast Gradient Sign Method (FGSM) [5] and Projected Gradient Descent (PGD) [6], construct adversarial perturbations by exploiting local gradient information of the loss function. Optimization-based approaches, including the Carlini and Wagner (C&W) attack [7] and the Jacobian-based Saliency Map Attack (JSMA) [8], formulate adversarial example generation as a constrained optimization problem. More recently, Generative Adversarial Network (GAN)-based methods [9] have attracted increasing attention by transforming adversarial example generation from an iterative optimization process into a feed-forward generation paradigm. Representative methods in this category include AdvGAN [10], AdvGAN++ [11], and LpAdvGAN [12]. Despite their effectiveness, most existing adversarial attack methods primarily focus on maximizing the attack success rate, often at the expense of imperceptibility. In particular, GAN-based approaches tend to generate perturbations that are diffusely distributed across the entire image, leading to visually noticeable artifacts. This limitation significantly restricts their applicability in real-world scenarios such as privacy protection and stealthy security testing, where imperceptibility is a critical requirement [13]. Therefore, achieving a favorable trade-off between attack effectiveness and visual imperceptibility remains an open challenge.

In our previous work, LpAdvGAN improves the attack success rate by identifying vulnerable pixel clusters and guiding the GAN training process accordingly. However, the perturbations generated by this strategy often lack sufficient visual subtlety, making the adversarial noise perceptible to human observers. This observation motivates a key question: *where should adversarial perturbations be injected to maximize attack effectiveness while preserving imperceptibility?*

To address this challenge, we propose **SpAdvGAN**, a sparse attention-guided adversarial example generation framework. SpAdvGAN introduces a local sparse attention extractor that explicitly restricts perturbations to semantically important regions of the image, enabling fine-grained control over perturbation distribution. By allowing only $n\sqrt{n}$ spatial positions to be attended, the proposed sparse attention mechanism avoids unnecessary global noise diffusion and preserves critical visual structures. Furthermore, we incorporate a dual autoencoder architecture within the GAN generator to facilitate structured noise fusion, capturing hierarchical feature representations at multiple levels. Additional training stabilizations, includ-

*Corresponding Author.

This work was supported by the Young Science and Technology Talent Support Program for Doctoral Students of Inner Mongolia Association for Science and Technology under Grant No. QTBS2518.

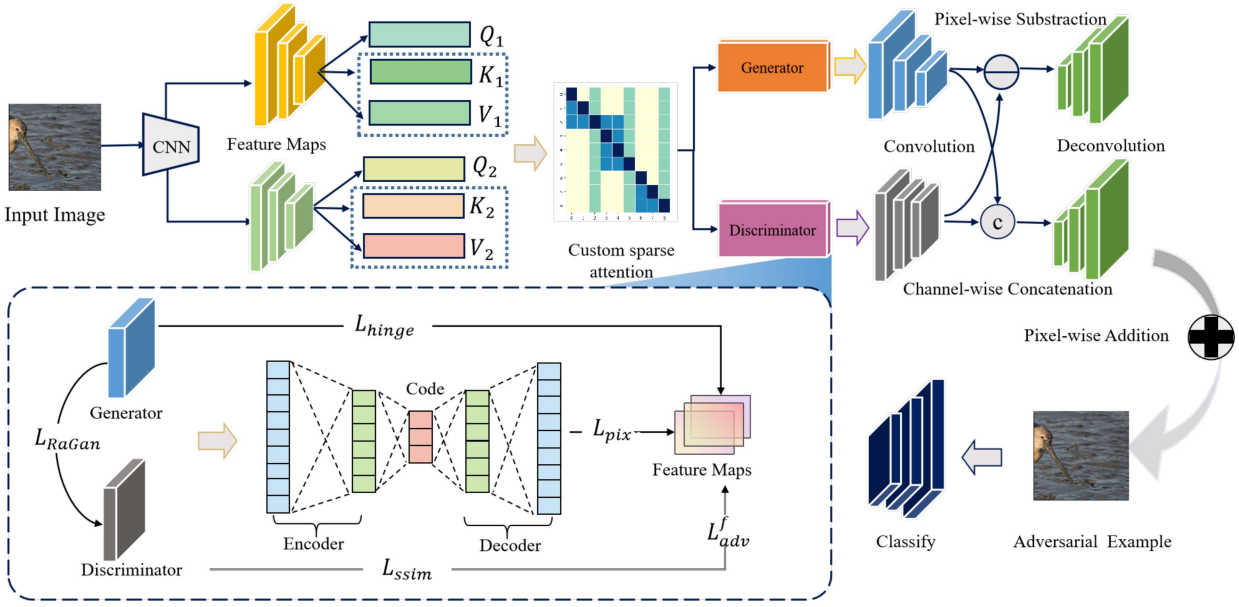

Fig. 1. Overall architecture of the proposed SpAdvGAN framework. The sparse semantic attention extractor identifies semantically important regions and constrains perturbations to a limited set of spatial locations. The extracted attention features guide the GAN-based generator with dual autoencoders to synthesize imperceptible adversarial perturbations under multiple perceptual constraints.

ing spectral normalization and a relative discriminator, are employed to further enhance adversarial generation without modifying the Lipschitz constraint.

The main contributions of this work are summarized as follows:

- We propose a sparse semantic attention-guided GAN framework that restricts adversarial perturbations to semantically salient regions, enabling fine-grained control over perturbation placement.
- We develop a dual autoencoder-based generator for structured noise fusion, improving imperceptibility without sacrificing attack effectiveness.
- We validate the proposed method through extensive semi-white-box experiments on multiple datasets and target models, demonstrating consistent improvements over existing adversarial attack methods.

II. Methodology

A. Problem Formulation

Let $x \in \mathbb{R}^{H\times W\times C}$ denote a clean input image with ground-truth label y, and let $f(\cdot)$ represent a target classification model. The objective of adversarial example generation is to construct a perturbed image

$$x' = x + \delta, \tag{1}$$

where δ denotes the adversarial perturbation, such that the perturbed example x' induces misclassification by the target model while remaining imperceptible to human observers.

Formally, under a semi-white-box attack setting, the adversarial example should satisfy

$$f(x') \neq y, \tag{2}$$

while minimizing the perceptual difference between x and x'. In this work, we aim to jointly optimize attack effectiveness and imperceptibility by constraining both the magnitude and the spatial distribution of adversarial perturbations.

B. Overview of SpAdvGAN

To achieve the above objective, we propose SpAdvGAN, a sparse attention-guided adversarial example generation framework. As illustrated in Fig. 1, SpAdvGAN is composed of three tightly coupled components that collaboratively generate imperceptible adversarial examples. Specifically, SpAdvGAN first employs a sparse semantic attention extractor to identify image regions with high semantic importance and restrict adversarial perturbations to a limited set of spatial locations. The extracted attention features are then used as conditional guidance for a GAN-based generator, which synthesizes adversarial perturbations under the constrained semantic structure. To further enhance imperceptibility, the generator integrates a noise fusion and loss constraint module that incorporates dual autoencoders and multiple perceptual loss terms, enabling fine-grained control over perturbation magnitude and visual consistency. Together, these components form a unified framework that balances attack effectiveness and visual imperceptibility.

C. Sparse Semantic Attention Extractor

To identify image regions with high semantic importance, we employ a self-attention mechanism due to its ability to model long-range dependencies through query-key-value (Q, K, V) interactions. However, conventional self-attention suffers from quadratic computational complexity with respect to the number of pixels. To address this issue, SpAdvGAN adopts a

sparse self-attention mechanism inspired by information flow sparsification [14].

Specifically, instead of attending to all spatial positions, the sparse attention extractor selects only $n\sqrt{n}$ positions from the feature map, resulting in a sparse attention matrix. This design significantly reduces computational overhead while preserving the most informative semantic relationships. The extracted sparse attention map serves as a semantic mask that constrains perturbations to a limited set of salient regions, preventing unnecessary noise diffusion across the entire image.

—

D. Noise Fusion Generator with Dual Autoencoders

The generator $G(\cdot)$ aims to produce adversarial perturbations that can effectively mislead the target classifier while maintaining visual fidelity. To better fuse sparse attention features and improve robustness, SpAdvGAN incorporates two parallel deep autoencoder branches within the generator.

Given an input image x, the two autoencoders encode the input into independent latent representations:

$$z_1 = E_1(x), \quad z_2 = E_2(x), \tag{3}$$

where $E_1(\cdot)$ and $E_2(\cdot)$ denote two encoder networks. The latent features are then fused through feature aggregation to generate the perturbation:

$$\delta = G(x) = F(z_1, z_2), \tag{4}$$

where $F(\cdot)$ denotes the feature fusion operator.

This dual-autoencoder design enables the generator to capture hierarchical feature representations at multiple levels, facilitating smoother noise synthesis and improved imperceptibility. Spectral normalization [15], [16] is applied to stabilize GAN training and prevent gradient explosion without explicitly enforcing a Lipschitz constraint.

E. Loss Function Design

To jointly optimize attack effectiveness and imperceptibility, SpAdvGAN employs a composite loss function that integrates adversarial supervision, perturbation magnitude control, perceptual consistency, and stable GAN training. Each loss term plays a complementary role in guiding the generator toward producing visually inconspicuous yet highly effective adversarial examples.

First, to encourage successful misclassification by the target classifier, we introduce an adversarial misclassification loss. Given a clean image x and its adversarial counterpart $x' = x + G(x)$, the generator is trained to maximize the discrepancy between the classifier prediction and the target label t:

$$L_{\text{adv}}^{f} = \mathbb{E}_x \left[\Gamma_f(x + G(x), t) \right], \tag{5}$$

where $\Gamma_f(\cdot, t)$ denotes the classification loss function of the target model.

To explicitly constrain the magnitude of adversarial perturbations, we incorporate a hinge loss that penalizes perturbations exceeding a predefined threshold:

$$L_{\text{hinge}} = \mathbb{E}_x \left[\max\left(0, \|G(x)\|_2 - c\right) \right], \tag{6}$$

where c controls the maximum allowable perturbation norm. This term prevents excessive noise injection and stabilizes the adversarial generation process.

In addition to global perturbation control, pixel-level consistency between the adversarial example and the original image is enforced through an L_2-based pixel loss:

$$L_{\text{pix}} = \|x' - x\|_2. \tag{7}$$

This constraint limits fine-grained pixel deviations and helps preserve local visual fidelity.

To further enhance perceptual similarity at the structural level, we introduce a structural similarity (SSIM) loss, which measures luminance, contrast, and structural consistency between x and x':

$$L_{\text{SSIM}}(x, x') = \frac{(2\mu_x\mu_{x'} + c_1)(2\sigma_{xx'} + c_2)}{(\mu_x^2 + \mu_{x'}^2 + c_1)(\sigma_x^2 + \sigma_{x'}^2 + c_2)}, \tag{8}$$

where μ and σ denote image means and variances, respectively, and c_1, c_2 are stability constants. This loss encourages the generator to preserve high-level structural information that is critical to human perception.

Finally, to stabilize GAN training and improve the realism of generated adversarial examples, we adopt a relativistic average GAN (RaGAN) loss [17]. Let $C(x)$ denote the discriminator output before activation. The RaGAN objective compares the realism of real samples x_r and generated samples x_f in a relative manner:

$$\begin{aligned} L_{\text{RaGAN}} = & -\mathbb{E}_{x_r}\left[\log\tanh\big(C(x_r) - \mathbb{E}_{x_f}C(x_f)\big)\right] \\ & -\mathbb{E}_{x_f}\left[\log\tanh(C(x_f) - \mathbb{E}_{x_r}C(x_r))\right]. \end{aligned} \tag{9}$$

This formulation enables the discriminator to focus on relative realism, thereby providing more informative gradients for the generator. Taken together, the above loss terms jointly guide the training of SpAdvGAN by balancing attack effectiveness, perturbation magnitude, perceptual consistency, and adversarial realism. Specifically, the adversarial misclassification loss encourages successful attacks against the target classifier, while the hinge loss explicitly constrains the overall perturbation magnitude. The pixel-level and SSIM losses collaboratively preserve fine-grained visual details and structural similarity between adversarial and clean examples. Meanwhile, the relativistic average GAN loss stabilizes adversarial training and improves the realism of generated samples.

The overall optimization objective of SpAdvGAN is formulated as:

$$L = L_{\text{adv}}^{f} + \alpha L_{\text{RaGAN}} + \beta L_{\text{hinge}} + \gamma L_{\text{pix}} + \omega L_{\text{SSIM}}, \tag{10}$$

where α, β, γ, and ω are weighting coefficients that control the relative importance of each loss component during training.

F. Training Strategy

During training, the discriminator and generator are optimized alternately. The generator aims to synthesize imperceptible adversarial examples that both mislead the classifier and resemble real images, while the discriminator learns to distinguish real samples from generated adversarial examples. Training proceeds until a Nash equilibrium is reached.

Algorithm 1 Training Procedure of SpAdvGAN

Require: Clean dataset $\mathcal{X}$, target classifier f, generator G, discriminator D

for each training iteration **do**
  Sample a mini-batch $x \sim \mathcal{X}$
  Extract sparse semantic attention features from x
  Generate perturbation $\delta = G(x)$
  Construct adversarial example $x' = x + \delta$
  Update discriminator D using L_{RaGAN}
  Update generator G using total loss L
end for
return Trained generator G

TABLE I
CLASSIFICATION ACCURACY OF TARGET MODELS ON CLEAN IMAGES.

Dataset	Model	Accuracy (%)
CIFAR-10	ResNet20	99.49
	ResNet32	99.52
	ResNet44	99.42
	DenseNet	85.91
ImageNet-Compatible	Inception v3	99.59

G. Algorithm Description

The overall training procedure of SpAdvGAN is summarized in Algorithm 1.

III. Experiments and Analysis

In this section, we evaluate the effectiveness of the proposed SpAdvGAN under a semi-white-box attack setting. The experimental study focuses on two key aspects: (1) attack effectiveness, measured by the attack success rate against different target models, and (2) imperceptibility of the generated adversarial examples, assessed using multiple perceptual quality metrics. All experiments are conducted on the CIFAR-10 and ImageNet-Compatible datasets, and SpAdvGAN is compared with representative gradient-based and GAN-based adversarial attack methods.

A. Datasets and Target Models

We conduct experiments on the CIFAR-10 dataset [18], which contains 60,000 RGB images from 10 classes with a resolution of 32×32. Following common practice in adversarial attack evaluation, 50,000 images are used for training the adversarial generator, and 5,000 correctly classified images are randomly selected from the test set for evaluation. The target models for CIFAR-10 include ResNet20, ResNet32, ResNet44, and DenseNet.

To further evaluate the generalization ability of SpAdvGAN in more realistic scenarios, we adopt the ImageNet-Compatible dataset introduced in the NIPS Adversarial Attack and Defense Competition [19]. This dataset consists of 1,000 RGB images with a resolution of 299×299, each correctly classified by the target model prior to attack. Inception v3 is used as the target classifier for ImageNet-Compatible experiments. The classification accuracy of all target models on clean images is reported in Table I.

TABLE II
TOP-1 CLASSIFICATION ACCURACY (%) OF TARGET MODELS UNDER ADVERSARIAL ATTACKS (LOWER IS BETTER).

Dataset	Model	FGSM	PGD	LpAdvGAN	SpAdvGAN
CIFAR-10	ResNet20	45.54	16.74	12.74	**10.70**
	ResNet32	35.12	13.79	13.67	**10.49**
	ResNet44	39.93	20.58	18.41	**11.74**
	DenseNet	27.14	29.59	13.48	**11.49**
ImageNet-Compatible	Inception v3	85.4	80.4	75.8	**70.1**

B. Semi-White-Box Attack Setting

In the semi-white-box setting, the adversarial generator is trained using a surrogate model with access to model architecture and prediction outputs, while the attack performance is evaluated on target models without access to their internal parameters. This setting reflects a practical threat model where attackers possess partial knowledge of the target system. All reported results are obtained by directly applying the trained SpAdvGAN generator to the target models without further optimization.

C. Implementation Details

SpAdvGAN is implemented using the PyTorch framework. The generator and discriminator are optimized using the Adam optimizer with a fixed learning rate of 0.001. For fair comparison, all baseline attack methods are implemented using the same datasets and evaluation protocols. The perturbation magnitude is constrained by an ℓ_∞ bound of 0.3 for CIFAR-10 and 8 for ImageNet-Compatible. The SpAdvGAN model is trained for 200 iterations under identical settings. All experiments are conducted on an NVIDIA A100 GPU using PyTorch 1.3.

D. Evaluation Metrics

To comprehensively evaluate imperceptibility, we adopt six commonly used metrics, including L_0, L_1, L_2, and L_∞ norms, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) [20]. Among them, PSNR and SSIM are treated as primary perceptual quality metrics, as they are closely aligned with human visual perception. Attack effectiveness is measured using the attack success rate (ASR), defined as the percentage of adversarial examples that successfully induce misclassification by the target model.

E. Attack Effectiveness Comparison

We compare SpAdvGAN with representative gradient-based attack methods, including FGSM [5] and PGD [6], as well as the GAN-based baseline LpAdvGAN [12]. Table II reports the top-1 classification accuracy of target models under adversarial attacks, where lower values indicate stronger attacks. As shown in Table II, SpAdvGAN consistently achieves lower classification accuracy across different target models, indicating higher attack success rates under the semi-white-box setting.

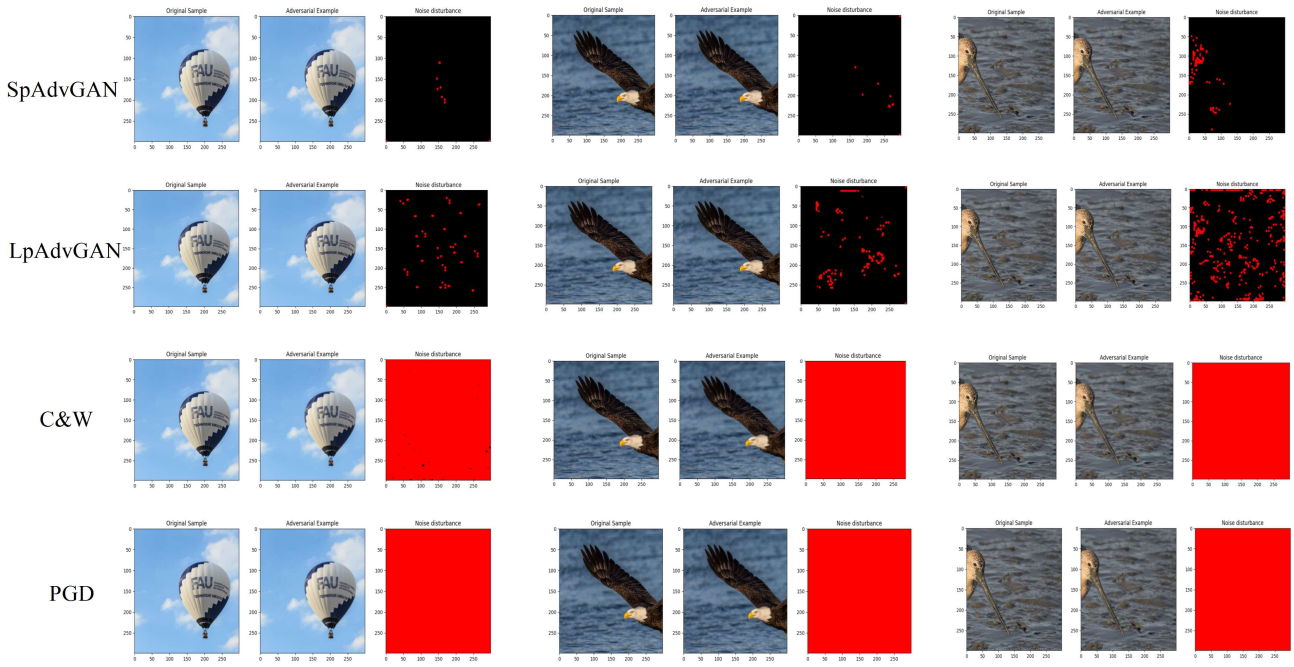

Fig. 2. Distribution of perturbations between the original examples and the adversarial examples generated by SpAdvGan.

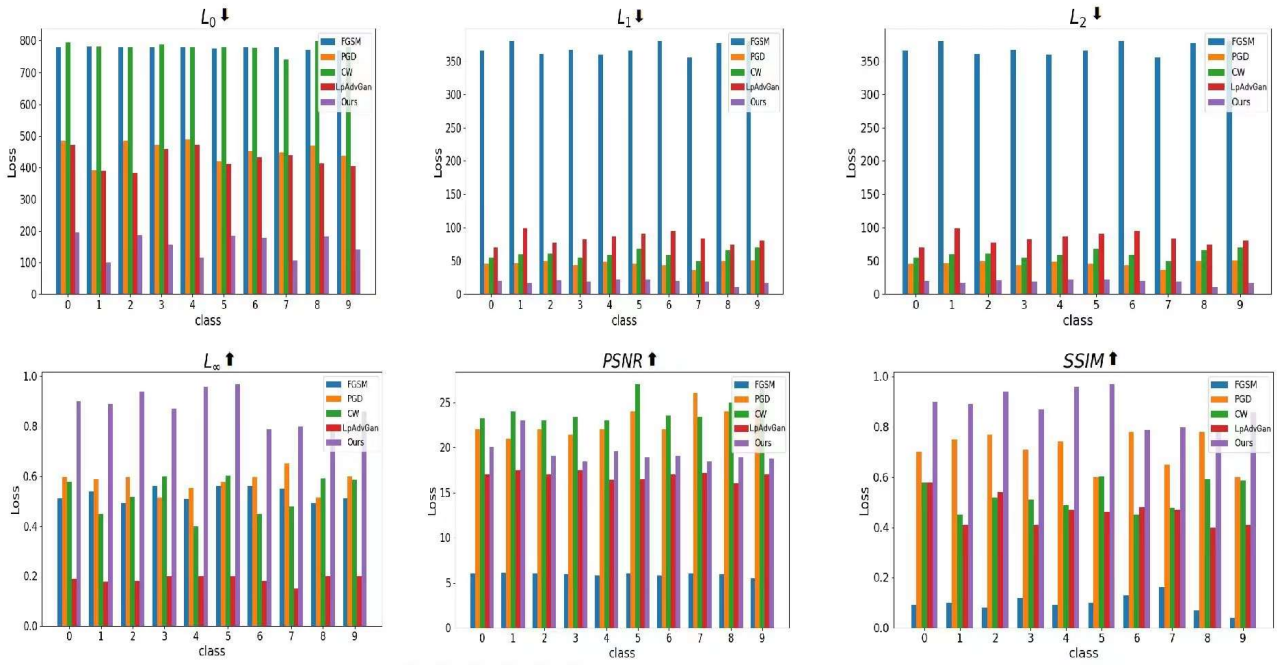

Fig. 3. SpAdvGan metric evaluation of selected adversarial examples based on 10 categories of the ImageNet-Compatible dataset and visualization of pixel differences between original and adversarial examples.

F. Ablation Study

We conduct an ablation study under the semi-white-box setting to analyze the contribution of each component in SpAdvGAN. Table III summarizes the results in terms of attack success rate (ASR) and imperceptibility measured by SSIM. Removing the sparse semantic attention (SSA) causes the most pronounced performance degradation, leading to noticeable drops in both ASR and SSIM. This indicates that constraining

TABLE III
ABLATION STUDY OF SPADVGAN COMPONENTS ON CIFAR-10.

Variant	S	D	L_s	L_p	R	N	ASR (%)	SSIM
Full	✓	✓	✓	✓	✓	✓	**89.3**	**0.912**
w/o SSA	×	✓	✓	✓	✓	✓	86.2	0.835
w/o DAE	✓	×	✓	✓	✓	✓	87.8	0.885
w/o L_{SSIM}	✓	✓	×	✓	✓	✓	89.2	0.862
w/o L_{pix}	✓	✓	✓	×	✓	✓	89.3	0.895
w/o RaGAN	✓	✓	✓	✓	×	✓	85.4	0.887
w/o SN	✓	✓	✓	✓	✓	×	84.1	0.892

perturbations to semantically salient regions is critical for achieving a favorable balance between attack effectiveness and imperceptibility. Without SSA, perturbations tend to spread across the image, resulting in inferior perceptual quality.

The dual autoencoder (DAE) also plays an important role. Its removal consistently reduces imperceptibility, suggesting that structured noise fusion is essential for generating visually coherent adversarial perturbations. Moreover, the perceptual loss terms exhibit complementary effects: removing the SSIM loss mainly affects structural similarity, while removing the pixel-level loss degrades fine-grained visual consistency. Finally, removing the relativistic discriminator (RaGAN) or spectral normalization (SN) leads to moderate but consistent performance degradation, reflecting their role in stabilizing GAN training. Overall, these results confirm that all components contribute synergistically to SpAdvGAN, with SSA and DAE being the most influential factors.

G. Imperceptibility Analysis

In addition to attack effectiveness, we analyze the imperceptibility of adversarial examples generated by different methods. Figure 2 visualizes the spatial distribution of perturbations produced by SpAdvGAN, demonstrating that perturbations are more sparsely distributed and aligned with semantically meaningful regions. Quantitative evaluation using perceptual metrics further confirms that SpAdvGAN achieves superior imperceptibility compared with baseline methods. Representative examples and metric comparisons on the ImageNet-Compatible dataset are shown in Fig. 3, where SpAdvGAN exhibits higher structural similarity and reduced visual artifacts. Overall, the experimental results indicate that SpAdvGAN achieves a favorable trade-off between attack effectiveness and imperceptibility. Compared with existing adversarial attack strategies, SpAdvGAN generates adversarial examples that are both more difficult to detect and more effective in misleading target models, making it a promising approach for evaluating model robustness in practical settings.

IV. CONCLUSION

In this paper, we propose SpAdvGAN, a sparse semantic attention-guided generative adversarial framework for adversarial example generation. By integrating a sparse attention mechanism with a dual autoencoder-based generator, SpAdvGAN enables fine-grained control over the spatial distribution and structure of adversarial perturbations. This design effectively improves the imperceptibility of adversarial examples while maintaining strong attack effectiveness. Overall, SpAdvGAN provides an effective and flexible approach for generating imperceptible adversarial examples and offers a practical tool for evaluating the robustness of deep neural networks in realistic attack scenarios.

REFERENCES

[1] J. Hayes and G. Danezis, "Learning universal adversarial perturbations with generative models," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 43–49.

[2] K. R. Mopuri, A. Ganeshan, and R. Venkatesh Babu, "Generalizable data-free objective for crafting universal adversarial perturbations," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 10, 2018, pp. 2452–2465.

[3] K. R. Mopuri *et al.*, "Nag: Network for adversary generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 742–751.

[4] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[7] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[8] T. Combey, A. Loison, M. Faucher, and H. Hajri, "Probabilistic jacobian-based saliency maps attacks," *Machine learning and knowledge extraction*, vol. 2, no. 4, pp. 558–578, 2020.

[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," vol. 35, no. 1, 2018, pp. 53–65.

[10] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," *arXiv preprint arXiv:1801.02610*, 2018.

[11] S. Jandial, P. Mangla, S. Varshney, and V. Balasubramanian, "Advgan++: Harnessing latent layers for adversary generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[12] D. Yang and J. Liu, "Lpadvgan: Noise optimization based adversarial network generation adversarial example," in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8.

[13] Z. Chen, Z. Wang, J.-J. Huang, W. Zhao, X. Liu, and D. Guan, "Imperceptible adversarial attack via invertible neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 414–424.

[14] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[15] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[16] F. Gogianu, T. Berariu, M. C. Rosca, C. Clopath, L. Busoniu, and R. Pascanu, "Spectral normalisation for deep reinforcement learning: an optimisation perspective," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3734–3744.

[17] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE nternational conference on computer vision*, 2017, pp. 2794–2802.

[18] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," vol. 25, 2012.

[20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.