ObjectAdv: Object-Level Unrestricted Adversarial Attacks via Diffusion Models

Shijie Zhao[1,3], Zhenyu Liang[2,4†], Xing Yang[2,3*], Haoqi Gao[2,3], Anjie Peng[1,3*], Hui Zeng[1]

[1]Southwest University of Science and Technology, Mianyang 621010, China
[2]Advanced Laser Technology Laboratory of Anhui Province, College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China
[3]Jianghuai Advanced Technology Center, Hefei 230037, China
[4]Information Security Research Center, Hefei Comprehensive National Science Center, Hefei 230037, China

zhaoshijie18@mails.swust.edu.cn, {liangzy21, yangxing17}@nudt.edu.cn,
{gaohaoqi77, penganjie200012}@163.com, zengh5@mail2.sysu.edu.cn

Abstract

Unrestricted adversarial attacks aim to fool DNNs by generating effective yet photorealistic examples. However, previous methods usually rely on global perturbations to enhance attack performance, which inevitably introduces visual distortions. To reduce visual distortions in the background, we propose a diffusion-based framework that focuses on local perturbations to generate object-level unrestricted adversarial examples (ObjectAdv). Since the cross-attention maps of Stable Diffusion contain the object information, we directly leverage the attention maps to localize the semantic region of object where for attacking. Second, a prompt-switching strategy is proposed for both imperceptibility and attack capacity. Specifically, to preserve layout and object shape of clean image, a prompt of true category is used at early denoising steps. At the later steps, we propose a well-designed prompt to guide the diffusion model to generate transferable adversarial examples. This local attack may cause inconsistency between the perturbed object and the background in adversarial examples. An FFT-based edge smoother is utilized to ensure seamless blending of the edges. ObjectAdv achieves an average ASR of 99.2% in white-box test on the ImageNet-compatible dataset, and outperforms existing methods on defense performance (+5%) and image quality metrics, e.g., SSIM of 0.9140 (+0.1048) and FID of 25.63 (-19.27).

Introduction

Deep neural networks (DNNs) are shown to be vulnerable to adversarial attacks (Madry et al. 2018; Dong et al. 2018), which pose serious security risks to practical applications. Most existing methods (Li et al. 2023; Wang et al. 2024; Ma et al. 2024; Zhu et al. 2024a) perturb images under the l_p-norm constraint, globally applying perturbations via various optimization approaches (Ge et al. 2023) to maximize attack performance. However, such globally distributed l_p-norm perturbations often induce visible artifacts that are easily perceived by human eyes.

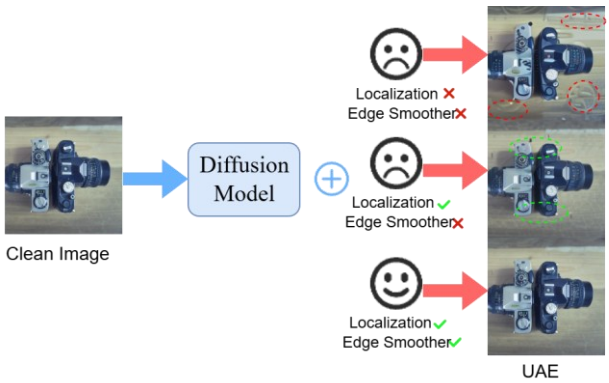

Figure 1: An UAE generated by our method. As seen, localization scheme erases shape distortions on background (red markers), and edge smoother alleviates the inconsistency between object and background (green markers).

To overcome the perceptual limitations of l_p-attacks, unrestricted adversarial attacks have been proposed to generate effective and natural-looking examples. Currently, unrestricted attacks fall into two categories. One is the semantic content-based attack, which modifies high-level semantic content of images, e.g., shape (Xiao et al. 2018), texture (Bhattad et al. 2019; Qiu et al. 2020), and color (Yuan et al. 2022), to generate unrestricted adversarial examples (UAEs). These methods generate photorealistic UAEs in most cases. However, for images with complicated texture, these methods exhibit significant texture or color changes (see Figure 2). Besides, since the semantic content-based UAEs are generated by CNN-based surrogate models,

[†] The co-first author Zhenyu Liang contributes equally to this work.
* Anjie Peng and Xing Yang are corresponding authors.

cAdv DiffAttack AdvDiff VENOM

Figure 2: UAEs (second row) generated by different methods. The first row shows clean images. Obviously, the UAE generated by cAdv exhibits unnatural colors. Other DM-based UAEs exhibit visual distortion.

the transferable performance of fooling Transformer models needs to be improved.

The other category is generative model-based attack. GAN-based attacks (He et al. 2022) may suffer from mode collapse and visual distortion issues. Recent methods integrate diffusion models (Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Rombach et al. 2022) into adversarial optimization processes to generate UAEs. They work in an image-editing-like manner, either repeatedly adding noise and denoising to eliminate noise patterns in final UAEs (Xue et al. 2023; Xie et al. 2024) or optimizing latent embeddings (Chen et al. 2024). With powerful energy of diffusion models (DM), DM-based methods achieve high transferability. However, added adversarial perturbations may destroy the distribution of latent, which causes visual distortions in UAEs (see Figure 2). Zhang et al. (Zhang et al. 2025) speculate that the high transferability of DM-based methods is mainly attributed to unreasonable visual distortions of UAEs, which violates photorealistic purpose of unrestricted attacks. AdvAD (Li et al. 2024a) achieves high imperceptibility, but with low transferability.

In a word, previous unrestricted attacks that craft global perturbations encounter the bottleneck to balance the attack efficacy and image quality. Intuitively, the local attack which adds adversarial perturbation in the part of clean image is a way of improving quality of UAE. Adversarial examples of local attacks (He et al. 2022; Ming et al. 2024) have already demonstrated satisfactory imperceptibility. The main issues of local attack are: (1) which region for attack and (2) how to improve attack capacity.

In this work, we propose a novel DM-based object-level unrestricted adversarial attack (ObjectAdv) to improve attack efficacy and imperceptibility. Inspired by watermark (Zhang et al. 2024), steganography (Yang et al. 2019) and local perturbation attacks (Zhu et al. 2024b) that pursue image quality, we select the object region which generally contains edges and complex texture for attacking. On the other hand, the object region plays an important role in the classification decision, and thus attacking such region benefits the disturbance of the classification results.

Considering that the denoising process is an innate coarse-to-fine synthesis (Choi et al., 2021; Chefer et al. 2023), the pipeline of progressive generation is adopted. In the early stage, a normal denoising without attack is executed, aiming at generating the layout and object shape of clean image in UAE. Then, we grant attack capacity to UAE via attacking the object region in the fine-detail stage. For a unified attack framework, an attention-based localization scheme is proposed. Specifically, the cross-attention maps of Stable Diffusion conditioned by the category prompt are utilized to localize the object region. Only the object region will be attacked, while other parts are kept as those of clean image. Figure 1 shows that the attention-based localization scheme erases the visual distortion at background. A prompt-switching strategy is proposed to switch a prompt of clean image category to a well-designed prompt of other categories in the later stage. This strategy generates UAEs maintaining layout and shape of clean image in the early stage, and gradually forces UAEs to leave the decision boundary of true category and approach that of the designed error category. The local attack may cause inconsistency between object and background. An FFT-based edge smoother is designed to produce a contextually coherent UAE, as shown in Figure 1.

Our contributions are summarized as follows:

- We first propose a unified DM-based framework to generate object-level unrestricted adversarial examples, integrating modules of attention-based localization, prompt guided attack and edge smoother. This framework combines the advantages of local attack for image quality and the powerful generation capacity of DMs, to solve the bottleneck problems of existing attacks in balancing attack efficacy and image quality.
- We propose a novel prompt-switching strategy to guide the diffusion model to generate UAEs that are imperceptible and transferable.
- Extensive experiments on a variety of model architectures and defense methods demonstrate that our attack achieves better attack performance and imperceptibility than previous local attacks and DM-based attacks.

Related Works

Semantic Content-based Unrestricted Attacks

Unrestricted adversarial attacks address the perceptual limitations of traditional l_p-attacks (Tran et al. 2022; Zhu et al. 2023). The l_p-attacks impose pixel-level perturbations, which tend to produce perceptible artifacts. In contrast, unrestricted attacks generate natural examples.

Early methods introduce perturbations by modifying

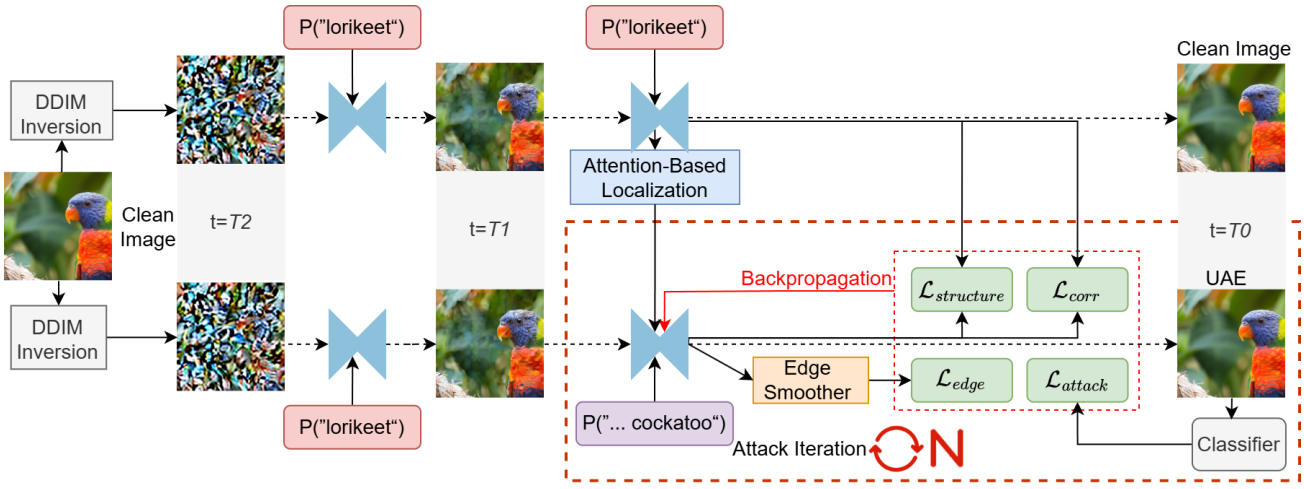

Figure 3: Overview of the ObjectAdv framework. $P(w)$ represents the prompts that guide the generation of UAE. In the stage ranging in $[T2, T1]$, a normal denoising without attack is executed. In $[T1, T0]$, we first utilize attention-based localization to locate the attacked object region. Then, we execute the attack by iteratively (N iterations) optimizing the composite loss in red dashed box. The orange dashed box denotes the whole attack process in $[T1, T0]$.

high-level semantic content, including shape, texture, and color. For example, stAdv (Xiao et al. 2018) and ADef (Alaifari, Alberti, and Gauksson 2019) iteratively introduce minor deformations to generate UAEs. cAdv (Bhattad et al. 2019) and SemanticAdv (Qiu et al. 2020) manipulate image texture or style to induce adversarial behavior. Color-based attacks, such as NCF (Yuan et al. 2022), Colorfool (Shamsabadi et al. 2020) and ACE (Zhao, Liu, and Larson 2020), adjust hue, saturation, or channels to improve visual naturalness. However, the UAEs generated by these methods exhibit texture or color changes. Besides, their attack performance falls behind that of l_p-attacks.

DM-based Unrestricted Attacks

Recent research leverages the latent space of DM to generate UAEs. These methods operate by repeatedly adding and removing noise to eliminate noise patterns in final UAEs or optimizing latent embeddings.

The pioneer DiffAttack (Chen et al. 2024) and ACA (Chen et al. 2023) combine the optimization of adversarial loss with stable diffusion to generate UAEs. Diff-PGD (Xue et al. 2023) and AdvDiffuser (Xie et al. 2024) incorporate famous PGD (Madry et al. 2018) into the diffusion steps, denoising the UAEs to improve robustness. AdvDiff (Dai, Liang, and Xiao 2024) proposes two interpretable adversarial guidance techniques to generate UAEs during denoising. VENOM (Zhang et al. 2025) unifies image content generation and adversarial synthesis into a single reverse diffusion process. Recent works, SCA (Pan et al. 2024), SemDiff (Dai et al. 2025), ScoreAdv (Huang and Tang 2025), APA (Jiang et al. 2025), and NatADiff (Collins et al. 2025), enhance the transferability and imperceptibility of UAEs by optimizing latent semantic boundaries and latent alignment.

With powerful generation capacity, DM-based attacks have demonstrated great potential in improving transfer performance. However, DM-based attacks may generate unexpected low-quality images. Because the adversarial perturbations added in the denoising process destroy the distribution of latent, which causes visual distortion in UAEs. In the following, we propose a DM-based local attack to generate imperceptible UAEs.

Method

Problem Formulation

Given a clean image x and its true label y, the optimization problem for an unrestricted adversarial attack against a deep learning classifier $F_\theta(\cdot)$ is formulated in Eq. 1.

$$\begin{aligned} &\arg\max_{x_{adv}} F_\theta\big(x_{adv} = \mathrm{DM}(z, w)\big) \neq y, \\ &\text{s.t. } x_{adv} \text{ is imperceptible with } x \end{aligned} \tag{1}$$

Under the diffusion model $\mathrm{DM}(\cdot)$, our task is to find the latent z and conditional prompt w to generate a deceptive yet imperceptible UAE x_{adv}.

Overall Framework

Our framework based on Stable Diffusion is shown in Figure 3. It first leverages DDIM inversion (Song, Meng, and Ermon 2020) to map the clean image into the diffusion latent space through reverse deterministic sampling (Eq. 2).

Figure 4: Mask results from attention-based localization.

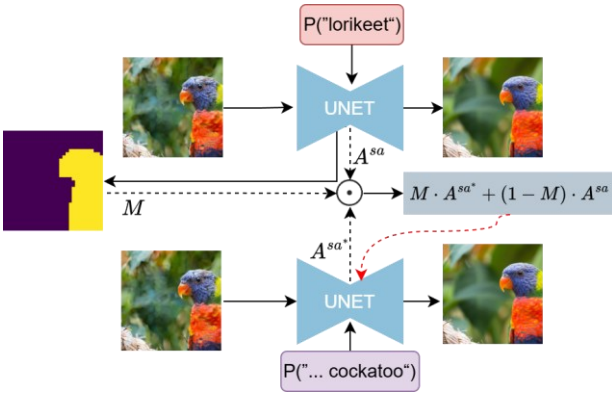

Figure 5: Attention-based localization. For simplicity, only the processing of M and A^{sa*} are shown here.

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(z_t) \quad (2)$$

In Eq.2, z_t denotes the latent representation of the clean image x, $\bar{\alpha}_t$ is cumulative noise scaling factor at time step t, $\epsilon_\theta(\cdot)$ is the diffusion neural network.

Under our framework, UAE is generated in the denoising process. In the whole process, the information of the clean image and prompt of category guide the generation of UAE. Since the denoising is an innate coarse-to-fine synthesis, our method divides the denoising into two stages $[T2, T1]$ and $[T1, T0]$. In $[T2, T1]$, a normal denoising without attack is executed, making a pure Gaussian noise become an UAE with similar appearance as the clean image. In $[T1, T0]$, we apply local attack on object region, and generate the final details of the UAE. Notice that, we adjust $T1$ to control the runtime. A larger $T1$ means a shorter attack stage $[T1, T0]$. ObjectAdv and its fast version ObjectAdv-I, which only use one denoising step are both considered. During the attack, an attention-based localization scheme, a prompt-switching strategy, and an edge smoother are designed. Details of these modules are as follows.

Attention-Based Localization

This scheme shown in Figure 5 is used in the stage $[T1, T0]$ to locate the object region for attacking. The attention maps of Stable Diffusion contain object information (Liu et al., 2024). The cross-attention maps are conditioned by the category prompt, so they are utilized to localize the object region. Notice that, for a unified framework, the widely used segmentation tools, e.g., SAM (Kirillov et al., 2023) are not utilized for location. Only the object region will be attacked, while other parts are kept as those of the clean image.

Figure 6: Results of prompt guidance at different stages (shown in colored bars). The second UAE shows a deformation that is generated by using a wrong prompt ("…cockatoo") in the whole stage $[T2, T1]$. The third image, guided by the correct prompt ("lorikeet"), obtains better image quality but loses attack ability. The fourth image, guided by our prompt-switching strategy, balances image quality and attack efficacy.

We first calculate cross-attention maps of clean image for each word of prompt as in Eq. (3). The prompt w is selected as the true category of clean image. $w[i]$ is the i^{th} word of prompt. $\mathrm{Average}(\mathrm{Cross}(\cdot))$ is accumulation and average of all cross-attention maps in $[T1, T0]$. An up-sampling operation is performed to match the size of obtained cross-attention maps $A^{ca(i)}$ with that of latent z_t.

$$A^{ca(i)} = \mathrm{Up}\left(\mathrm{Average}\left(\mathrm{Cross}(z_t, t, w[i]; \mathrm{DM})\right)\right) \quad (3)$$

Next, Eq. (4) is executed to obtain the mask matrix $M^{(i)}$. $Norm(\cdot)$ is a normalization operation. For multi-word categories (e.g., "monarch butterfly"), we execute Eq. (5) and Eq. (6) to obtain the final mask M and cross-attention maps A^{ca}. In Eq. (5), "|" is an or operation which expands the object region, n is the number of words in the prompt. Figure 4 shows some mask results.

$$M^{(i)}(k,j) = \begin{cases} 1, & Norm(A^{ca(i)}(k,j)) > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

$$M = M^{(1)} | M^{(2)} \ldots | M^{(n)} \quad (5)$$

$$A^{ca} = \frac{1}{n}\sum_{i=1}^{n} A^{ca(i)} \quad (6)$$

After getting the localization mask M, only the object region will be attacked, while other parts are preserved as those of the clean image to improve imperceptibility. The preserve operation is done as in Eq. (7), where z^*, A^{ca*}, and A^{sa*} are the latent, cross-attention maps, and self-attention maps of the UAE, respectively.

$$\begin{aligned} z^* &= M \cdot z^* + (1-\mathrm{M}) \cdot z \\ A^{ca*} &= M \cdot A^{ca*} + (1-\mathrm{M}) \cdot A^{ca} \\ A^{sa*} &= M \cdot A^{sa*} + (1-\mathrm{M}) \cdot A^{sa} \end{aligned} \quad (7)$$

Models	Attacks	TIME(s)↓	ASR(%)↑	SSIM↑	PSNR↑	LPIPS↓	FID↓	HyperIQA↑	TReS↑
Res-50	cAdv	18.3	97.3	0.8290	24.91	0.1197	49.52	0.5419	69.38
	NCF	10.4	89.9	0.3175	14.73	0.3089	82.59	0.5201	67.60
	ACA	125.3	90.0	0.5137	18.09	0.3361	45.60	0.5533	73.89
	DiffAttack	29.9	96.3	0.7665	22.66	0.1370	62.60	0.5658	75.57
	Di-Mask	28.4	92.1	0.7801	23.84	0.1265	58.49	0.5792	76.12
	AdvDiff	20.5	98.8	0.2658	12.84	0.6983	43.23	0.5129	68.92
	VENOM	9.3	98.7	0.6788	20.31	0.1856	62.20	0.5524	74.89
	ObjectAdv	15.2	**99.2**	**0.9143**	**28.08**	**0.0532**	**26.79**	**0.6099**	**78.05**
	ObjectAdv-*I*	**4.5**	**100.0**	**0.8876**	**26.61**	**0.0665**	**29.53**	**0.5723**	**77.74**
VGG-19	cAdv	17.4	96.7	0.7929	23.68	0.1427	50.69	0.5350	68.65
	NCF	9.9	92.9	0.3171	14.77	0.3104	78.81	0.5215	68.04
	ACA	119.0	92.4	0.6541	18.21	0.3309	45.84	0.5531	73.89
	DiffAttack	28.4	97.1	0.7573	22.25	0.1236	63.90	0.5830	75.83
	Di-Mask	27.8	94.6	0.7657	23.44	0.1205	57.90	0.5897	76.14
	AdvDiff	19.5	98.9	0.2523	12.21	0.6645	41.57	0.4948	67.25
	VENOM	8.8	98.5	0.6616	18.94	0.1947	64.50	0.5681	75.47
	ObjectAdv	14.4	**99.5**	**0.9135**	**28.00**	**0.0537**	**26.31**	**0.6096**	**78.07**
	ObjectAdv-*I*	**4.2**	**100.0**	**0.8953**	**25.42**	**0.0582**	**30.52**	**0.5764**	**76.98**
ViT-B	cAdv	19.2	97.7	0.8092	24.50	0.1330	49.47	0.5376	68.82
	NCF	10.9	72.6	0.3425	15.06	0.2914	67.67	0.5232	67.76
	ACA	131.6	81.2	0.6492	18.07	0.3395	44.81	0.5580	74.45
	DiffAttack	31.4	95.3	0.7570	22.25	0.1248	65.40	0.5589	75.35
	Di-Mask	30.5	85.2	0.7628	23.13	0.1165	57.93	0.5628	75.98
	AdvDiff	21.5	98.4	0.2795	13.47	0.7321	44.90	0.4912	67.60
	VENOM	9.8	96.5	0.6523	19.93	0.1948	63.40	0.5353	74.25
	ObjectAdv	16.0	**99.0**	**0.9140**	**28.06**	**0.0538**	**25.63**	**0.6088**	**77.74**
	ObjectAdv-*I*	**4.8**	**100.0**	**0.8930**	**25.25**	**0.0615**	**29.45**	**0.5792**	**75.38**

Table 1: Results of untargeted white-box attack. The best result is **bolded**, and the second-best result is underlined.

Prompt-Switching Strategy

Prompt is a key parameter of image generation in DM. Since the denoising is a coarse-to-fine synthesis, a prompt-switching strategy is proposed to enhance attack capacity of UAEs. To preserve layout and object shape of clean image, in $[T2, T1]$, we use a prompt of true category. To enhance attack capacity, a well-designed prompt is used in $[T1, T0]$. We use the category that is a synonym to the true category of clean image as prompt. ChatGPT-4 is used to find synonym prompt of true category. Figure 6 shows that the proposed strategy achieves satisfactory performance.

Optimization Design

Attack Classifier. Following (Chen et al. 2024), the loss function $\mathcal{L}_{attack}$ with the optimization variable z^* is designed to maximize the cross-entropy loss $J(\cdot)$, forcing the classifier F_θ to predict an error label. In Eq. (8), w^* is a well-designed prompt, and $\mathrm{Denoise}(\cdot)$ is the denoising process ranging in $[T1, T0]$. "×" is Hadamard product.

$$\mathcal{L}_{attack} = -J(\mathrm{DM}(z^* \times M + z \times (1-M), w^*), y; F_\theta),$$
$$\text{where } z^* = z_0^* = \underbrace{\mathrm{Denoise} \circ \cdots \circ \mathrm{Denoise}}_{T1 \to T0}(z_t^*) \quad (8)$$

Align with Error Prompt. To enhance transferability of UAEs, we align the image content with the well-designed prompt w^* which expresses an error category of the clean image. This alignment will force the semantic content of UAE to be far away from the true category, thus enhancing transferability. Since the prompt acts directly on the cross-attention maps, we calculate the loss $\mathcal{L}_{corr}$ from the cross-attention maps A^{ca} of clean image, and A^{ca*} that of adversarial image by Eq. 9.

$$\mathcal{L}_{corr} = \rho\big(\mathrm{Average}(A^{ca}), \mathrm{Average}(A^{ca*})\big) \quad (9)$$

Here, ρ is Pearson correlation coefficient.

Structure Preservation. To preserve content structure of clean image, self-attention maps A^{sa*} of UAE are forced to approximate A^{sa} of the clean image in a l_2-norm manner by Eq. 10.

$$\mathcal{L}_{structure} = \| A^{sa} - A^{sa*} \|_2^2 \quad (10)$$

Attacks	Res-50 *	VGG-19	Mob-v2	ViT-B	Swin-B	DeiT-B	Avg.
cAdv	78.6/70.5/97.3	28.1/24.3/36.9	33.6/33.0/43.7	21.8/19.0/28.2	15.5/13.7/21.4	19.4/17.1/25.1	32.8/29.1/42.1
NCF	1.6/0.5/89.9	1.1/0.3/72.2	1.0/0.4/71.7	0.3/0.0/37.9	0.3/0.0/26.8	0.3/0.0/29.3	0.8/0.2/54.6
ACA	12.5/80.2/90.0	8.4/55.2/68.3	8.3/53.2/70.0	5.2/36.5/51.9	5.8/36.8/51.3	6.0/37.2/52.0	7.7/49.9/63.9
DiffAttack	54.2/42.0/96.3	40.3/30.6/**75.6**	41.5/31.6/**77.1**	27.8/21.6/51.2	30.0/22.8/56.2	26.3/20.7/50.1	36.7/28.2/67.8
Di-Mask	71.7/52.5/92.1	40.5/29.2/47.3	40.8/29.7/48.9	19.9/13.7/27.1	21.5/14.3/29.9	20.7/14.7/30.2	35.8/25.7/45.9
AdvDiff	1.2/85.2/98.8	0.8/52.2/65.8	0.7/51.8/69.5	0.0/38.2/**57.9**	0.0/38.6/**58.5**	0.0/39.1/**60.6**	0.5/50.9/**68.5**
VENOM	23.2/48.3/98.7	21.0/34.1/62.5	20.0/33.3/63.2	11.7/20.5/41.2	14.7/24.8/45.7	12.2/21.7/40.6	17.1/30.1/58.7
ObjectAdv	**98.5/99.2/99.2**	**53.2/55.5**/55.5	**53.8/56.0**/56.0	**30.1**/36.8/36.8	**34.2/40.8**/40.8	**32.5**/37.4/37.4	**50.4/54.3**/54.3
Object…-*I*	**99.8/100/100**	**57.4/59.1**/59.1	**59.3/63.3**/63.3	**34.3/42.1**/42.1	**38.7/45.2**/45.2	**34.8/42.6**/42.6	**54.1/58.7**/58.7

Table 2: ASR results of untargeted white-box (denoted by *) and black-box attacks, where the first/second/third results are calculated from UAEs with PSNR >22/FID<45/without image quality selection.

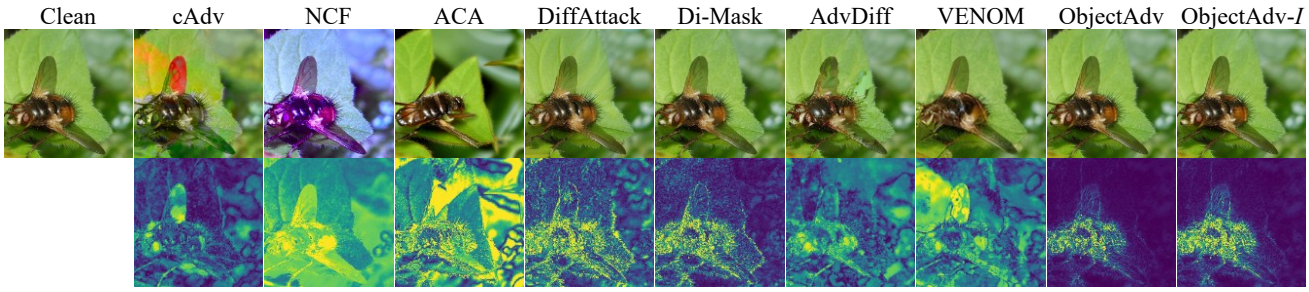

Figure 7: Visualization of UAEs generated on Res-50 and the corresponding amplified perturbations (×20).

FFT-based Edge Smoother. Inspired by Li et al. (Li et al. 2024b), we design an FFT-based edge smoother for a contextually coherent UAE. The initial M in Eq. 5 is dilated to obtain M_d, and then dilated layers are formed as in Eq. 11.

$$I_{L,d} = A^{sa*} \times M_d, I_{G,d} = A^{sa} \times M_d \quad (11)$$

In the frequency domain, a 2D Fourier transform $\mathcal{F}(\cdot)$, a low-pass filter H, and an inverse transform $\mathcal{F}^{-1}(\cdot)$, are applied as in Eq. 12.

$$ds = H \times \left(\mathcal{F}(I_{L,d}) - \mathcal{F}(I_{G,d})\right), M_{\text{dm}} = |\mathcal{F}^{-1}(ds)| \quad (12)$$

Through a composition of binarization and morphological closing and filling functions, M_{dm} is operated to obtain M'. The edge loss is defined in Eq. 13.

$$\mathcal{L}_{edge} = |M' \times (A^{sa} - A^{sa*})| \quad (13)$$

Final Loss. The final loss is shown in Eq. 14, where α, β, γ and δ represent the weight factors of each loss.

$$\arg\min_{x_{adv}} \mathcal{L} = \alpha\mathcal{L}_{attack} + \beta\mathcal{L}_{corr} + \gamma\mathcal{L}_{structure} + \delta\mathcal{L}_{edge} \quad (14)$$

Experiments

Experimental Settings

Datasets and Models. The widely used ImageNet-compatible Dataset (Kurakin et al. 2018), consisting of 1,000 images from ImageNet's validation set (Deng et al. 2012), is selected. Following (Chen et al. 2024), three convolutional neural networks (CNNs), including Res-50 (He et al. 2016), VGG-19 (Simonyan and Zisserman 2014), and Mob-v2 (Sandler et al. 2018), three vision transformers (ViTs), including ViT-B (Dosovitskiy et al. 2021), Swin-B (Liu et al. 2021), and DeiT-B (Touvron et al. 2021), and three defense models, including HGD (Liao et al. 2018), R&P (Xie et al. 2018), and DiffPure (Nie et al. 2022), are selected as the target models for attack.

Attack Methods. Unrestricted attacks including cAdv (Bhattad et al. 2019) and NCF (Yuan et al. 2022), as well as DM-based attacks, including DiffAttack and its local version Di-Mask (Chen et al. 2024), ACA (Chen et al. 2023), AdvDiff (Dai et al. 2024), and VENOM (Zhang et al. 2025), are compared.

Evaluation Metrics. To evaluate imperceptibility, three reference-based metrics, including SSIM (Wang et al. 2004), PSNR (Zhang et al. 2022), and LPIPS (Huynh-Thu and Ghanbari 2008) and three non-reference metrics, including

Attacks	HGD	R&P	DiffPure	Avg.
cAdv	22.2	20.6	32.1	25.0
NCF	29.8	28.2	36.5	31.5
ACA	31.3	30.1	38.9	33.4
DiffAttack	37.9	36.7	50.5	41.7
Di-Mask	38.4	32.8	52.9	41.6
AdvDiff	45.4	46.8	53.2	48.5
VENOM	39.4	39.7	55.2	44.8
ObjectAdv	**50.3**	**50.5**	**60.5**	**53.8**
ObjectAdv-I	**53.8**	**54.2**	**65.1**	**57.4**

Table 3: ASR (%) comparisons under defense schemes.

FID (Heusel et al. 2017), HyperIQA (Su et al. 2020), and TReS (Golestaneh et al. 2022) are adopted.

Implementation Details. DDIM full sampling is set to $T = 50$ steps (5 inversion steps), attack iterations are set to $\mathrm{N} = 20$. ObjectAdv and ObjectAdv-I are executed at the interval $[T1, T0] = [45, 50]$ and $[T1, T0] = [49, 50]$, respectively. The weight factors α, β, γ and δ in Eq. 14 are set to 10, 1, 100 and 100, respectively.

Attack Performance Comparison

White-Box Attacks. Table 1 reports the untargeted attack performance and imperceptibility of nine methods across three attacked models. For white-box attacks, our method achieves the best results in terms of ASR, imperceptibility metrics and run time. Our method obtains ASR of [99.2% 100%] when attacking different models.

For imperceptibility, local attacks perform better than global attacks. Our method adopts structure preservation and edge smoother to further improve image quality, and obtain much better scores. For the attack on ViT-B, compared to the second-best method, ObjectAdv achieves SSIM of 0.9140 (+0.1048), and FID of 25.63 (-19.27). Notice that ObjectAdv-I costs the least run time, but achieves nearly the best ASR or image quality scores. Figure 7 visualizes the perturbation distribution of UAEs.

Transferability & Robustness

Due to lack of explicit image quality constraints, these unrestricted attacks may generate some unnatural or distorted UAEs, which dramatically violate the semantic content of clean images. We argue that those poor-quality UAEs are not suitable for testing attack performance. To select natural-looking UAEs for testing, we set a threshold of PSNR or FID as the criterion.

Table 2 shows that our method achieves the best transferability on high-quality UAEs. Surprisingly, for the compared methods, there are big gaps among the ASRs with/without image quality selection, indicating that the high transferability of these methods is mainly attributed to unreasonable visual distortions. Table 3 reports the robust performance of UAEs under defense models.

loca-tion	Prompt-Switching	Edge Smoother	ASR(%)↑	FID↓
	√		99.7	42.27
√		√	93.8	24.83
√	√		99.4	30.46
√	√	√	99.2	26.79

Table 4: Ablation study of the key modules.

	M-N	M-1	M-2	M-3	M-4
ASR(%)↑	99.5	99.2	90.3	96.2	84.6
FID↓	39.56	26.79	29.98	30.44	34.50

Table 5: The results of attacking Res-50 by different masks, where "M-N" means without using mask (i.e., global attack), "M-1" is our method, "M-2" and "M-3" obtain the mask via Di-Mask and SAM, respectively, "M-4" adopts a random mask.

Ablation Study

Table 4 reports white-box results of the key modules of ObjectAdv attacking Res-50. As seen, only using the prompt switching strategy obtains the highest ASR, but deteriorates image quality. The attention-based localization scheme and edge smoother mainly contribute to the image quality, where the union of them obtains the best FID score.

The mask is an important module, because it decides which object region for attacking. Table 5 shows that the proposed mask M-1 achieves the best results, owing to the fact that the attention-based localization scheme finds a proper local region highly related to the object for attacking. Table 5 also shows that M-N, i.e., global attack, obtains the worst FID score, indicating that the local attack is a better choice to improve the image quality of UAE.

Conclusion

This paper proposes a novel object-level unrestricted adversarial attack framework, combining the advantage of local attack for image quality and the powerful generation capacity of Stable Diffusion model, to solve the bottleneck problems of existing attacks at balancing attack efficacy and image quality. This framework is able to generate effective and imperceptible UAEs. Experiments demonstrate that our attack achieves better attack performance and imperceptibility than those of previous local attacks and DM-based attacks. The object-level attacks armed with diffusion model provide new ideas for generating effective and imperceptible unrestricted adversarial examples. In the future, we will explore how to utilize DM to further improve the transferability of unrestricted adversarial examples.

Acknowledgments

This work was partially supported by Hefei Comprehen-sive National Science Cen-ter (KY23C503) and the Youth Independent Innovation Science Fund Projects of National University of Defense Technology (ZK24-47).

References

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

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.

Li, J.; He, Z.; Luo, A.; Hu, J.-F.; Wang, Z. J.; and Kang, X. 2024a. AdvAD: Exploring Non-Parametric Diffusion for Imperceptible Adversarial Attacks. *Advances in Neural Information Processing Systems 37*, 52323–52353.

Wang, K.; He, X.; Wang, W.; and Wang, X. 2024. Boosting Adversarial Transferability by Block Shuffle and Rotation. In *Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24336–24346.

Ma, A.; Farahmand, A.-M.; Pan, Y.; Torr, P.; and Gu, J. 2024. Improving Adversarial Transferability via Model Alignment. In *European Conference on Computer Vision*, 74–92. Cham: Springer Nature Switzerland.

Alaifari, R.; Alberti, G. S.; and Gauksson, T. 2019. ADef: An Iterative Algorithm to Construct Adversarial Deformations. In *7th International Conference on Learning Representations*.

Xiao, C.; Zhu, J.-Y.; Li, B.; He, W.; Liu, M.; and Song, D. 2018. Spatially Transformed Adversarial Examples. *In 6th International Conference on Learning Representations*.

Qiu, H.; Xiao, C.; Yang, L.; Yan, X.; Lee, H.; and Li, B. 2020. SemanticAdv: Generating Adversarial Examples via Attribute-Conditioned Image Editing. In *European Conference on Computer Vision*, 19–37. Cham: Springer International Publishing.

Bhattad, A.; Chong, M. J.; Liang, K.; Li, B.; and Forsyth, D. A. 2019. Unrestricted Adversarial Examples via Semantic Manipulation. In *International Conference on Learning Representations*.

Yuan, S.; Zhang, Q.; Gao, L.; Cheng, Y.; and Song, J. 2022. Natural Color Fool: Towards Boosting Black-Box Unrestricted Attacks. *Advances in Neural Information Processing Systems 35*, 7546–7560.

Zhao, Z.; Liu, Z.; and Larson, M. 2020. Adversarial Color Enhancement: Generating Unrestricted Adversarial Images by Optimizing a Color Filter. In *31st British Machine Vision Conference 2020*. BMVA Press.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 1–20.

Dhariwal, P. and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *Advances in Neural Information Processing Systems 34*, 8780–8794.

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*, 10684–10695.

Xue, H.; Araujo, A.; Hu, B.; and Chen, Y. 2023. Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability. *Advances in Neural Information Processing Systems 36*, 2894–2921.

Xie, Y.; Guo, X.; Wang, C.; Liu, K.; and Chen, L. 2024. AdvDiffuser: Generating Adversarial Safety-Critical Driving Scenarios via Guided Diffusion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9983–9989.

Chen, J.; Chen, H.; Chen, K.; Zhang, Y.; Zou, Z.; and Shi, Z. 2024. Diffusion Models for Imperceptible and Transferable Adversarial Attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, Z.; Li, B.; Wu, S.; Jiang, K.; Ding, S.; and Zhang, W. 2023. Content-Based Unrestricted Adversarial Attack. *Advances in Neural Information Processing Systems 36*, 51719–51733.

Dai, X.; Liang, K.; and Xiao, B. 2024. AdvDiff: Generating Unrestricted Adversarial Examples Using Diffusion Models. In *European Conference on Computer Vision*, 93–109. Cham: Springer Nature Switzerland.

Zhang, H.; Chen, H.; and Zhao, G. 2025. VENOM: Text-Driven Unrestricted Adversarial Example Generation with Diffusion Models. *arXiv preprint arXiv:2501.07922.*

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14367–14376.

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.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2018. Mitigating Adversarial Effects Through Randomization. In *International Conference on Learning Representations*.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning.*

Liu, B.; Wang, C.; Cao, T.; Jia, K.; and Huang, J. 2024. Towards Understanding Cross and Self-Attention in Stable Diffusion for Text-Guided Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7817–7826.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial Examples in the Physical World. In *Artificial Intelligence Safety and Security*, 99–112. Chapman and Hall/CRC.

Deng, J.; Berg, A.; Satheesh, S.; Su, H.; Khosla, A.; and Fei-Fei, L. 2012. ImageNet Large Scale Visual Recognition Competition 2012 (ILSVRC2012). *See net.org/challenges/LSVRC.*

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Simonyan, K. and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556.*

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M. et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In I*nternational Conference on Learning Representations.*

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.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training Data-Efficient Image Transformers & Distillation Through Attention. In *International Conference on Machine Learning*, 10347–10357.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. I*EEE Transactions on Image Processing 13(4)*, 600–612.

Zhang, J.; Wu, W.; Huang, J.-T.; Huang, Y.; Wang, W.; Su, Y.; and Lyu, M. R. 2022. Improving Adversarial Transferability via Neuron Attribution-Based Attacks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14993–15002.

Huynh-Thu, Q. and Ghanbari, M. 2008. Scope of Validity of PSNR in Image/Video Quality Assessment. *Electronics Letters 44(13)*, 800–801.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems 30*.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3667–3676.

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 2022. No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1220–1230.

Shamsabadi, A. S.; Sanchez-Matilla, R.; and Cavallaro, A. 2020. Colorfool: Semantic Adversarial Colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1151–1160.

Pan, Z. H.; Chen, L. F.; Wu, W. B.; Cao, Y. H.; and Zheng, Z. B. 2024. SCA: Improve Semantic Consistent in Unrestricted Adversarial Attacks via DDPM Inversion. *arXiv preprint arXiv:2410.02240.*

Dai, Z. Y.; Liu, S. C.; He, R.; Wu, J. H.; Lu, N.; Fan, W. Q.; Li, Q.; and Tang, K. 2025. SemDiff: Generating Natural Unrestricted Adversarial Examples via Semantic Attributes Optimization in Diffusion Models. *arXiv preprint arXiv:2504.11923.*

Huang, C. H., and Tang, H. 2025. ScoreAdv: Score-based Targeted Generation of Natural Adversarial Examples via Diffusion Models. *arXiv preprint arXiv:2507.06078.*

Jiang, K. X.; Chen, Z. Y.; Guo, H. J.; Li, J. L.; Fu, J. Y.; Guo, P. X.; Tang, H.; Li, B.; and Zhang, W. Q. 2025. Enhancing Diffusion-based Unrestricted Adversarial Attacks via Adversary Preferences Alignment. *arXiv preprint arXiv:2506.01511.*

Collins, M.; Vice, J.; French, T.; and Mian, A. 2025. NatADiff: Adversarial Boundary Guidance for Natural Adversarial Diffusion. *arXiv preprint arXiv:2505.20934.*

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; et al. 2023. Segment Anything. In *IEEE/CVF International Conference on Computer Vision*.

Li, Q.; Guo, Y.; Zuo, W.; and Chen, H. 2023. Improving Adversarial Transferability via Intermediate-Level Perturbation Decay. *Advances in Neural Information Processing Systems 36*.

Zhu, R.; Zhang, Z.; Liang, S.; Liu, Z.; and Xu, C. 2024a. Learning to Transform Dynamically for Better Adversarial Transferability. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24273–24283.

Ge, Z.; Liu, H.; Wang, X.; Shang, F.; and Liu, Y. 2023. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems 36,* 70141–70161.

Ming, M.; Di, D.; Ren, P.; Wang, Y.; and Feng, X. 2024. Transferable structural sparse adversarial attack via exact group sparsity training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24696–24705.

Tran, H.; Lu, D.; and Zhang, G. 2022. Exploiting the Local Parabolic Landscapes of Adversarial Losses to Accelerate Black-Box Adversarial Attack. In *European Conference on Computer Vision,* 317–334. Cham: Springer Nature Switzerland.

Zhu, H.; Ren, Y.; Sui, X.; Yang, L.; and Jiang, W. 2023. Boosting Adversarial Transferability via Gradient Relevance Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4741–4750.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. In *ACM Transactions on Graphics*, 1–10.

Patashnik, O.; Garibi, D.; Azuri, I.; Averbuch-Elor, H.; and Cohen-Or, D. 2023. Localizing Object-Level Shape Variations with Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23051–23061.

Zhu, H.; Ren, Y.; Liu, C.; Sui, X.; and Zhang, L. 2024b. Frequency-Based Methods for Improving the Imperceptibility and Transferability of Adversarial Examples. *Applied Soft Computing 150*, 111088.

He, Z.; Wang, W.; Dong, J.; and Tan, T. 2022. Transferable sparse adversarial attack. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, S.; Zeng, B.; Feng, Y.; Gao, S.; Liu, X.; Liu, J.; Li, L.; et al. 2024b. Zone: Zero-Shot Instruction-Guided Local Editing. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6254–6263.

Zhang, L.; Liu, X.; Martin, A. V.; Bearfield, C. X.; Brun, Y.; and Guan, H. 2024. Attack-resilient image watermarking using stable diffusion. *Advances in Neural Information Processing Systems* 37, 38480–38507.

Yang, J.; Ruan, D.; Huang, J.; Kang, X.; and Shi, Y.-Q. 2019. An Embedding Cost Learning Framework Using GAN. *IEEE Transactions on Information Forensics and Security*, 839–851.