

Harnessing Global-Local Collaborative Adversarial Perturbation for Anti-Customization

Long Xu¹, Jiakai Wang^{2*}, Haojie Hao¹, Haotong Qin³, Jiejie Zhao², Xianglong Liu^{1,2}

¹Beihang University, ²Zhongguancun Laboratory, ³ETH Zürich

{xulonguid, haojiehao, xlliu}@buaa.edu.cn, {wangjk, zhaojiejie}@zgclab.edu.cn
 haotong.qin@pbl.ee.ethz.ch

Abstract

Though achieving significant success in personalized image synthesis, Latent Diffusion Models (LDMs) pose substantial social risks caused by unauthorized misuse (e.g., face theft). To counter these threats, the Anti-Customization (AC) method that exploits adversarial perturbations was proposed. Unfortunately, existing AC methods show insufficient defense ability due to the ignorance of hierarchical characteristics, i.e., global feature correlations and local facial attributes, leading to weak resistance to concept transfer and semantic theft in customization methods. To address these limitations, we are motivated to propose a **Global-Local Collaborated Anti-Customization (GoodAC)** framework to generate powerful adversarial perturbations by disturbing both feature correlations and facial attributes. To enhance the ability to resist concept transfer, we disrupt the spatial correlation of perceptual features that form the basis of model generation at a global level, thereby creating highly concept-transfer-resistant adversarial camouflage. To improve the ability to resist semantic theft, leveraging the fact that facial attributes are personalized, we designed a personalized and precise facial attribute distortion strategy locally, focusing the attack on the individual's image structure to generate strong camouflage. Extensive experiments on various customization methods, including Dreambooth and LoRA, have strongly demonstrated that our GoodAC outperforms other state-of-the-art approaches by large margins, e.g., over 50% improvements on ISM.¹

1. Introduction

The Latent Diffusion Model (LDM), which excels at generating highly realistic visual content, has garnered significant attention in recent years within the academic community [7, 12, 14, 25, 43], recognized by many researchers

*Corresponding author

¹Codes can be found at <https://github.com/xl-yaoyi/GoodAC>.

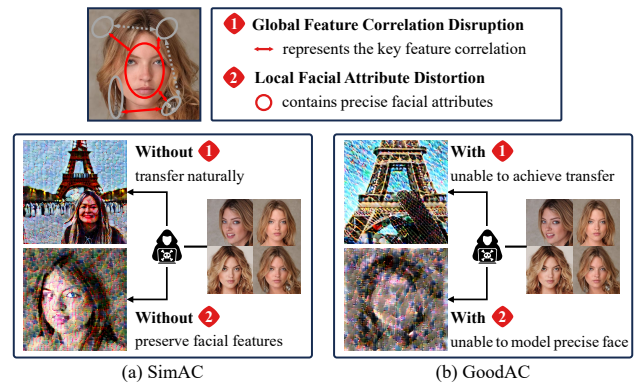


Figure 1. Comparison of SimAC and GoodAC. Our method further boosts its ability to de-identity.

as an emerging and efficient theoretical framework. The open-source nature of the Stable Diffusion models enables users to thoroughly understand and explore the innovative potential of LDM [28, 42]. Users can build personalized LDMs with just a few images depicting a specific concept (e.g., a particular person) and a unique identifier (e.g., “a sks person”). By finely adjusting the unique identifier, the model can efficiently produce high-quality images showcasing special themes, i.e., image customization [9].

However, as the saying goes, *every coin has two sides*, these innovative techniques in image generation also bring potential social risks, involving personal rights such as portrait information and reputation. Hackers, with easy access to publicly available data, can effortlessly collect a few images from social media profiles of public figures, often without permission, and employ them to generate false yet remarkably believable images, as seen in the case of Dreambooth [27]. Once created, these images can be maliciously posted online, potentially causing public relations crises or even political upheavals. Such nefarious activities will severely undermine the credibility of online information and pose serious threats to social safety [39]. These

emerging threats have spurred a response from the academic community, with some researchers advocating anti-customization (AC) approaches to protect image information from unauthorized use. These approaches use adversarial perturbations [10, 22, 37, 40] to protect image content, effectively prevent personal information from being stolen.

In detail, Glaze [29] is a pioneering approach to protect image information by adding perturbations. Recently, the SimAC [35] has advanced AC research to a new level. Though achieving results, existing studies still show some limitations, which can be summarized as: ❶ Weak resistance to concept transfer. Customization methods utilize perceptual features to learn the spatial correlations between concepts and scenarios. However, current AC methods only optimize based on the overall intermediate features of the model, fail to disrupt these spatial correlations and cannot effectively prevent customization methods from transferring concepts to new scenarios. ❷ Insufficient resistance to semantic theft. Different from naive adversarial attack, AC has given more missions to protect fine-grained semantic information, *e.g.*, face attribute details, from the theft of customization methods. Unfortunately, existing AC research has only preliminarily explored the role of high-frequency image features in gradient optimization, lacking comprehensive semantic segmentation and recognition, thereby resulting in insufficient resistance to semantic leakage.

To address the aforementioned issues, we propose a **Global-Local Collaborated Anti-Customization (GoodAC)** framework that generates robust adversarial perturbations by collaboratively disrupting global perceptual correlations and distorting the local face attributes. Regarding the resistance to concept transfer, perceptual features form the foundation for model recognition and generation [32]. The model relies on the spatial consistency of perceptual features to preserve core concept characteristics during concept transfer [38]. Thus, we prevent the model from accurately identifying and transferring key concept by disrupting the perceptual feature correlations globally. Specifically, we iteratively perform block angle transformation to continuously disrupt the perceptual feature correlations at the global level, thereby generating a camouflage against concept transfer. To resist the semantic theft, through detailed observation of existing research, we found that current studies do not fully utilize the fine-grained semantic attribute information of protected images, treating them as generic image perturbation tasks. Considering that facial attributes are personalized [44], we introduce a precise facial attribute distorting approach that targets to the local feature inside face feature space. Specifically, we precisely extract facial semantic attributes through detailed text descriptions and edge detection, thereby focusing the limited attack pixels on the image structure of individual subjects and perturbing more critical features to enhance protective performance.

Our main contributions are as follows:

- We are the first to identify the critical resistance that relates to the concept transfer and semantic theft play important role in anti-customization methods.
- We propose a global-local collaborated AC framework that simultaneously disrupts global perceptual feature correlations and local precise facial attributes, enhancing resistance to both the concept transfer and semantic theft.
- Extensive experiments on multiple Customization methods, including DreamBooth and LoRA, demonstrate the effectiveness of our method. Our approach reduces face similarity metrics (ISM) to below 10%, significantly outperforming state-of-the-art methods and providing better privacy protection across different training strategies.

2. Related Work

2.1. Customization of Latent Diffusion Models

Latent Diffusion Models (LDMs) [26, 34] have become a cornerstone of generative modeling [25]. Built upon a progressive denoising architecture, LDMs integrate Variational Auto-Encoders (VAEs) [16] to shift the diffusion process from image space to latent space, effectively mitigating overfitting at the pixel level [30] and significantly improving generation efficiency. Moreover, by incorporating cross-attention mechanisms and the CLIP encoder [24], LDMs demonstrate superior cross-modal generation capabilities, surpassing Conditional GANs (CGANs) [20] in synthesizing high-fidelity images from textual descriptions with greater accuracy and flexibility.

Customization refers to leveraging the powerful generative capabilities of LDMs to mimic given reference images representing specific concepts and generate new images of these concepts in different scenarios. In order to customize LDMs for specific concepts, methods like DreamBooth [27] and LoRA [13] are widely used. DreamBooth [27] fine-tunes the entire model using a few images and a unique identifier to associate new concepts with text embeddings. It modifies the original loss $\mathcal{L}_{cond}$ to $\mathcal{L}_{db}$ as follows:

$$\begin{aligned}\mathcal{L}_{cond}(\theta, x_0) &= \mathbb{E}_{x_0, t, c, \epsilon \in \mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(x_{t+1}, t, c)\|_2^2, \\ \mathcal{L}_{db}(\theta, x_0) &= \mathbb{E}_{x_0, t, t'} \|\epsilon - \epsilon_\theta(x_{t+1}, t, c)\|_2^2 \\ &\quad + \lambda \|\epsilon' - \epsilon_\theta(x'_{t'+1}, t', c_{pr})\|_2^2,\end{aligned}\tag{1}$$

where $\epsilon, \epsilon' \sim \mathcal{N}(0, I)$, $x'_{t'+1}$ is a noisy version of the class example x generated from the original LDM with prior prompt c_{pr} , and λ balances the importance of the prior term. LoRA (Low-Rank Adaptation) [13] freezes the original model weights and fine-tunes low-rank matrices inserted into the attention layers, reducing computational complexity and overfitting risks compared to full fine-tuning.

2.2. Anti-customization Studies

To prevent misuse of private images in customized LDMs, recent studies have explored adversarial techniques for privacy protection [17, 18, 33, 35, 41]. By introducing subtle perturbations to images, these methods aim to hinder the model’s ability to learn and transfer representations of target images. Early works like Glaze [29] applied style transformations to protect knowledge in latent diffusion models, while AdvDM [18] utilized the model’s inherent representations for attacks. Anti-Dreambooth [33] incorporated Projected Gradient Descent (PGD) [19] into latent diffusion model attacks, using substitute models to calculate adversarial perturbations and significantly improving attack success rates. More recent methods like SimAC [35] have leveraged differences in perturbation effectiveness at different timesteps, employing greedy algorithms to select optimal perturbations and further enhance attack success.

However, these methods primarily focus on improving optimization efficiency, but overlook two critical factors in customization: new scenarios and fine-grained facial attributes [18, 33, 35]. This oversight limits the ability to capture the contribution of hierarchical features to customization quality, resulting in suboptimal protection performance as well as limited transferability [36].

3. Approach

3.1. Problem Definitions

Let $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ denote the initial set of images that a user needs to protect. We define a perturbation function δ that maps the initial set of images $\mathcal{X}$ to the protected set $\mathcal{X}' = \{x_1 + \delta(x_1), x_2 + \delta(x_2), \dots, x_n + \delta(x_n)\}$. An adversary may acquire $\mathcal{X}$ and utilize it to fine-tune a noise prediction network ϵ_θ of the LDM under the DreamBooth [27] paradigm, so that the LDM can generate images containing the user’s information in different scenarios (i.e. concept transfer), and have clear user facial features and other details (i.e. semantic theft). Our objective is to design an optimal δ^* that minimizes the customization capability of the LDM under the trained parameters θ^* :

$$\begin{aligned} \delta^* &= \arg \min_{\delta} \mathcal{A}(\epsilon_{\theta^*}, \mathcal{X}'), \\ \text{s.t. } \theta^* &= \arg \min_{\theta} \sum_{x \in \mathcal{X}'} \mathcal{L}_{db}(\theta, x), \\ \text{and } \|\delta(x)\|_{\infty} &\leq \eta, \quad \forall x \in \mathcal{X}'. \end{aligned} \quad (2)$$

Here $\mathcal{L}_{db}$ is the loss function of DreamBooth [27] defined in Eq.(1), η is the constraint on perturbation budget, $\|\cdot\|_{\infty}$ is the L_{∞} norm, and $\mathcal{A}(\epsilon_{\theta^*}, \mathcal{X}')$ is an evaluation metric designed to assess the fidelity of images generated by the LDM with respect to the original images $\mathcal{X}$. This metric quantitatively reflects the model’s customization capability and will be elaborated in the experimental section.

3.2. Framework Overview

To generate adversarial camouflages with strong resistance to concept transfer and semantic theft, we propose a **Global-local collaborated Anti-Customization (GoodAC)** framework. This framework aims to generate robust adversarial perturbations by collaboratively disrupting global perceptual correlations and distorting the local face attributes. The overall framework is illustrated in Figure 2.

To enhance resistance to concept transfer, inspired by existing studies, perceptual features serve as the foundation for image generation [32], and LDM heavily relies on spatial consistency in perceptual features to maintain identity during concept transfer. Motivated by this insight, we propose to strengthen the robustness of protected images against concept transfer by globally disrupting spatial correlations within the perceptual feature space. Specifically, we extract feature maps from both the first down-sampling block and the last up-sampling block of the UNet, and employ the block angle transformation to disrupt spatial connectivity within these feature maps. Based on this, we design a loss function to optimize adversarial perturbations, effectively preventing the concept in protected images from being successfully transferred into novel scenarios.

In terms of resisting semantic theft, considering that fine-grained semantic features such as facial characteristics are crucial for protection, it is essential to accurately identify these details. This allows us to apply stronger perturbations locally to critical image regions, thereby weakening the ability of LDM to extract detailed user-specific information from the images. Specifically, we employ text inversion module to obtain precise semantic descriptions as semantic guidance, combined with edge detection techniques to accurately extract facial details. Consequently, perturbations are effectively focused on these critical regions, limiting the model’s capability to capture personalized and precise facial attributes, thereby increasing the difficulty of accurately reproducing details from the original images.

3.3. Global Feature Correlation Disruption

Perceptual features refer to critical feature representations extracted by deep learning models from specific objects during image processing. These features typically emerge within intermediate layers of the model, such as convolutional layers and residual blocks. Existing research has shown that perceptual features serve as the fundamental basis for both recognition and generation tasks [32]. During customization, the Latent Diffusion Model leverages spatial correlations within these perceptual features to effectively transfer personalized subject concepts into new scenarios for image generation. Inspired by this mechanism, we propose disrupting the spatial correlations within perceptual features to strengthen the resistance of anti-customization examples against such concept transfer.

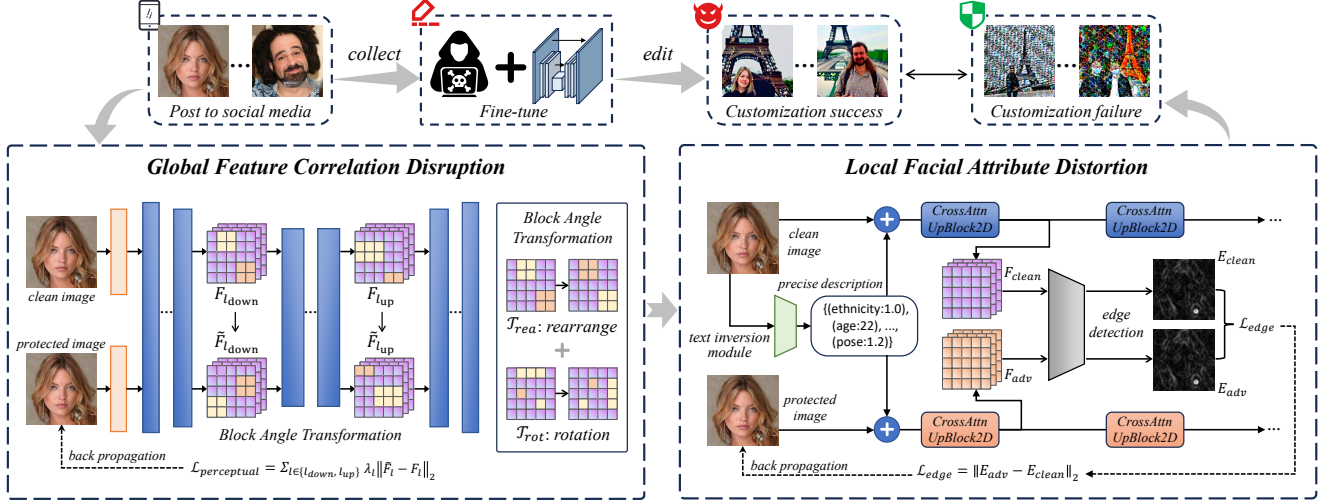


Figure 2. The overview of GoodAC. we first use global feature correlation disruption strategy to disrupt the spatial correlation of perceptual features. Then we use local facial attribute distortion strategy to attack precise individual facial attributes to generate strong camouflage.

Specifically, when LDM transfers concepts bound to the identifier “a *sk*s person” into novel scenarios (e.g., “a *sk*s person looking at mirror”), it relies on a set of spatially ordered perceptual features to maintain conceptual integrity. If we disrupt the spatial structure of these features, the model loses part of the stable information associated with the concept during generation, making it difficult to preserve the original characteristics of the “a *sk*s person”, thereby achieving an anti-customization effect.

To disrupt the spatial correlations among perceptual features, we introduce a feature map transformation called block angle transformation. On one hand, we partition the feature map F_l into blocks and rearrange them, thereby disrupting the internal continuity and spatial correlations of the feature map, making it difficult for the model to capture associative patterns, this procedure is defined as $\mathcal{T}_{rea}$. On the other hand, angle transformations such as rotation applied to the feature map F_l alter its global spatial arrangement, thus impairing the model’s understanding of orientation and shape, this procedure is defined as $\mathcal{T}_{rot}$. The complete process is formulated as follows:

$$\tilde{F}_l = \mathcal{T}_{rot}(\mathcal{T}_{rea}(F_l)), \quad l \in \{l_{down}, l_{up}\}, \quad (3)$$

where F_l and $\tilde{F}_l$ represent the feature maps from layer l for all channel before and after transformation respectively, and $l \in \{l_{down}, l_{up}\}$ indicates that the feature maps are extracted from the first down-sampling block and the last up-sampling block of the UNet. By calculating the differences in feature maps before and after the transformation, we construct a loss function $\mathcal{L}_{perceptual}$ to guide adversarial perturbations in more effectively interfering with perceptual

features. The definition of $\mathcal{L}_{perceptual}$ is as follows:

$$\mathcal{L}_{perceptual} = \sum_{l \in \{l_{down}, l_{up}\}} \lambda_l \|\tilde{F}_l - F_l\|_2, \quad (4)$$

where λ_l is the weighting coefficient for each layer l . $\|\cdot\|_2$ is the Frobenius norm, measuring L_2 difference between transformed and original feature maps. Based on the block-angle transformation strategy, we calculate the differences in feature maps before and after transformation, thereby generating adversarial camouflage that effectively enhances anti-customization effectiveness at a global level.

3.4. Local Facial Attribute Distortion

To enhance the resistance of anti-customization examples against semantic theft, existing AC research [35] has primarily focused on gradient-based optimization approaches, investigating the role of high-frequency image features in accelerating the optimization process. However, these methods have not effectively segmented or identified semantically meaningful information within images for targeted optimization, thereby exhibiting limited capability in mitigating semantic theft [2, 4, 8, 23]. Therefore, we aim to precisely locate and distort local facial attributes, thereby disrupting the LDM’s ability to learn fine-grained semantic information such as facial features, and ultimately enhancing the effectiveness of anti-customization.

Specifically, we propose Local Facial Attribute Distortion that leverages adaptive semantic guidance and edge detection to target critical features within the anti-customization perturbations, consisting of two key steps: (1) generating more detailed descriptions of the protected images to guide the LDM in encoding richer internal feature of fine-grained semantic attributes; and (2) applying

edge detection between feature maps of adversarially camouflaged and clean images during the perturbation process, and optimizing the perturbations by minimizing the L_2 loss between the detected edges.

We first employ a text inversion module² to generate fine-grained image descriptions, which serve as semantic guidance to enrich the internal features of the LDM. For image x to be protected, we generate semantic prompts $\mathcal{P} = \text{TextInversion}(x)$ which describe user attributes as a set of values (e.g., $\mathcal{P} = \{\text{ethnicity:1.0}, (\text{age:22}), \dots, (\text{pose:1.2})\}$). We select the top 15 attributes with the highest confidence scores and combine them into a refined prompt $\mathcal{S}$, guiding the LDM to generate image features with accurately facial attributes and enhance the precision of local disruption:

$$F_{\text{clean}} = l_{\text{up}}(x, \mathcal{S}), \quad F_{\text{adv}} = l_{\text{up}}(x_{\text{adv}}, \mathcal{S}). \quad (5)$$

Here, we extract features from the last up-sampling block of the UNet, denoted as l_{up} , and apply detector $\mathcal{C}$ to identify important local features:

$$E_{\text{clean}} = \mathcal{C}(F_{\text{clean}}), \quad E_{\text{adv}} = \mathcal{C}(F_{\text{adv}}), \quad (6)$$

where E_{clean} and E_{adv} represent the edge maps, highlighting the critical local attributes essential for customization. Based on this, we define the loss $\mathcal{L}_{\text{edge}}$ to guide the adversarial perturbation to selectively disrupt these key features:

$$\mathcal{L}_{\text{edge}} = \|E_{\text{adv}} - E_{\text{clean}}\|_2. \quad (7)$$

In summary, our method leverages precise semantic attributes and edge-focused optimization to target key image features, effectively impairing the model’s ability to extract sensitive facial information and providing robust protection.

3.5. Overall Training Process

We combine the global feature correlation disruption loss $\mathcal{L}_{\text{perceptual}}$ and the local facial attribute distortion loss $\mathcal{L}_{\text{edge}}$, with the contribution of $\mathcal{L}_{\text{edge}}$ controlled by a weighting factor β , to obtain the overall loss $\mathcal{L}_{\text{GoodAC}}$:

$$\mathcal{L}_{\text{GoodAC}} = \mathcal{L}_{\text{perceptual}} + \beta \cdot \mathcal{L}_{\text{edge}}. \quad (8)$$

Considering the improvement of SimAC in perturbation optimization efficiency, we incorporate the whole feature loss $\mathcal{L}_{\text{feat}}$ of SimAC into the overall optimization process, which is defines it as follows:

$$\mathcal{L}_{\text{feat}} = \mathbb{E} \|F_{l,t,\text{adv}} - F_{l,t,\text{clean}}\|_2, \quad (9)$$

where $F_{l,t,\text{clean}}$ and $F_{l,t,\text{adv}}$ denote the feature maps of the clean image and the perturbed image at specific time steps t and layers l . The selection of t and l follows the strategy proposed in SimAC.

²We implement text inversion module using [Image Prompt API](#).

Finally, we jointly optimize the GoodAC loss $\mathcal{L}_{\text{GoodAC}}$, the SimAC loss $\mathcal{L}_{\text{feat}}$, and LDM denoising loss $\mathcal{L}_{\text{cond}}$, to generate the anti-customization adversarial camouflage δ :

$$\begin{aligned} \delta &= \alpha \cdot \text{sign}(\nabla_{x_{\text{adv}}} \mathcal{L}_{\text{cond}} + \mathcal{L}_{\text{GoodAC}} + \gamma \cdot \mathcal{L}_{\text{feat}}), \\ \text{s.t. } x_{\text{adv}} &= x_{\text{adv}} + \delta, \quad \|\delta\|_{\infty} \leq \eta, \end{aligned} \quad (10)$$

where x_{adv} is the perturbed image to be protected, α is the step size for perturbation, $\text{sign}(\cdot)$ is a sign function, and γ is the weighting factor that controls $\mathcal{L}_{\text{feat}}$. Empirically, we set γ to 0.2. The overall training algorithm can be found in the supplementary material.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets

Following the settings established in previous studies [33, 35], we conduct experiments using two facial image datasets: CelebA-HQ [15] and VGGFace2 [3]. These datasets consist of images from approximately 50 individuals, with 15 high-quality facial images available per individual for customization purposes. The images of each individual are divided into three groups: A, B, and C, each containing five portrait images. Group A is used as clean reference facial images for model training, while group B represents the target images for generating anti-customization examples. Group C is not utilized in the experiments.

4.1.2. Models

Given that the open-source Stable Diffusion is the most popular LDM implementation in the community, we follow the experimental setup used in prior studies [33, 35] and conduct model training and generate anti-customization examples on SD-v2.1 [1], and quantitatively evaluate the effectiveness of anti-customization on SD-v2.1. Furthermore, to assess the performance of our method in a black-box setting, we qualitatively examine the effectiveness of the anti-customization examples generated on SD-v2.1 when applied to SD-v1.4. The corresponding results are provided in the supplementary material.

4.1.3. Baselines

For the customization method, we adopt DreamBooth [27] as the baseline due to its strong performance. For comparison in anti-customization, we select three state-of-the-art open-source methods: AdvDM [18], Anti-DreamBooth [33] (denoted as Anti-DB), and SimAC [35], all of which utilize adversarial attacks to protect user images from potential misuse caused by LDM-based customization. In addition, we qualitatively evaluate the transferability of anti-customization examples generated against DreamBooth when applied to LoRA [13], and the corresponding results are provided in supplementary material.

Table 1. CelebA-HQ Dataset Results: Comparison with other open-sourced anti-customization methods on CelebA-HQ. We evaluate the performance under four different prompts during customization. The underlined method is the baseline for measuring metric changes.

Method	“a photo of <i>sks</i> person”				“a dsrlr portrait of <i>sks</i> person”			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
AdvDM[18]	0.32	70.48	38.17	0.20	0.25	65.37	37.86	0.41
Anti-DB[33]	0.28	77.28	37.43	0.20	0.19	86.80	38.90	0.27
SimAC[35]	0.31	87.07	38.86	0.21	0.12	96.87	42.10	0.15
Ours	0.03 _{89.28%↓}	96.67 _{25.09%↑}	39.75 _{6.20%↑}	0.02 _{90%↓}	0.01 _{96%↓}	99.99 _{15.09%↑}	43.21 _{11.07%↑}	0.01 _{96.29%↓}
Method	“a photo of <i>sks</i> person looking at the mirror”				“a photo of <i>sks</i> person in front of eiffel tower”			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
AdvDM[18]	0.29	35.10	36.46	0.36	0.09	38.10	36.02	0.30
Anti-DB[33]	0.22	42.86	40.34	0.28	0.06	56.26	41.35	0.22
SimAC[35]	0.12	91.90	43.97	0.06	0.04	71.87	43.99	0.12
Ours	0.02 _{90.9%↓}	96.67 _{114.41%↑}	44.11 _{10.08%↑}	0.01 _{96.42%↓}	0.02 _{66.66%↓}	96.67 _{71.82%↑}	44.99 _{8.8%↑}	0.02 _{90.9%↓}

Table 2. VGGFace2 Dataset Results: Comparison with other open-sourced anti-customization methods on VGGFace2. We evaluate the performance under four different prompts during customization. The underlined method is the baseline for measuring metric changes.

Method	“a photo of <i>sks</i> person”				“a dsrlr portrait of <i>sks</i> person”			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
AdvDM[18]	0.32	63.87	38.89	0.20	0.31	68.03	37.42	0.39
Anti-DB[33]	0.31	65.28	38.03	0.22	0.23	77.86	38.88	0.28
SimAC[35]	0.31	80.17	40.61	0.22	0.11	96.33	41.10	0.12
Ours	0.03 _{90.32%↓}	93.33 _{42.97%↑}	39.31 _{3.37%↑}	0.04 _{81.82%↓}	0.04 _{82.61%↓}	96.67 _{24.16%↑}	43.97 _{13.09%↑}	0.02 _{92.86%↓}
Method	“a photo of <i>sks</i> person looking at the mirror”				“a photo of <i>sks</i> person in front of eiffel tower”			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
AdvDM[18]	0.29	38.32	39.88	0.29	0.20	39.10	36.02	0.35
Anti-DB[33]	0.24	46.56	39.31	0.26	0.11	53.56	40.32	0.24
SimAC[35]	0.11	90.90	44.66	0.06	0.09	60.34	42.05	0.11
Ours	0.02 _{91.67%↓}	96.52 _{107.30%↑}	46.30 _{17.78%↑}	0.02 _{92.31%↓}	0.05 _{54.54%↓}	76.66 _{43.13%↑}	42.52 _{5.46%↑}	0.09 _{62.5%↓}

4.1.4. Evaluation Metrics

We adopt four evaluation metrics to assess the effectiveness of anti-customization. Specifically, we consider (1) Face Detection Failure Rate (FDFR), which quantifies whether a face can be detected in the generated image; (2) Identity Score Matching (ISM), which measures identity similarity between generated and original images; (3) SER-FQA, a facial image quality assessment metric; and (4) BRISQUE, a widely used general-purpose no-reference image quality metric. The detailed computation procedures for these metrics are provided in the supplementary material.

4.1.5. Detailed Experimental Settings

We incorporate AC examples generation into DreamBooth training, with 50 steps and 6 PGD iterations per step. The perturbation budget η is constrained to 16/255, and the step size α is set to $5e^{-3}$. During testing, the DreamBooth model is fine-tuned on the generated AC examples with a batch size of 2, a learning rate of $5e^{-7}$, and 1000 training steps. Unless otherwise specified, the training and prior prompts

are “a photo of *sks* person” and “a photo of person”. All experiments are conducted on a single NVIDIA L20 GPU.

4.2. Anti-Customization Performance

In this section, we evaluate the anti-customization performance of our GoodAC under CelebA-HQ and VGGFace2 datasets. We first apply GoodAC and several other AC methods to generate adversarial examples on SD-v2.1, which are then used to fine-tune SD-v2.1 via DreamBooth with the default prompt “a photo of *sks* person”. We then perform inference using this prompt and three additional scenario-specific prompts to transfer the concept of “*sks* person” to various scenarios. For each prompt, we randomly select 32 generated images to compute all evaluation metrics and report their averages. As shown in Table 1 and Table 2, we draw following conclusions:

(1) The adversarial camouflage generated by GoodAC achieves significantly superior performance across both datasets. On CelebA-HQ, the maximum ISM reduction exceeds 50%, and FDFR surpasses 95%. On VGGFace2, ISM

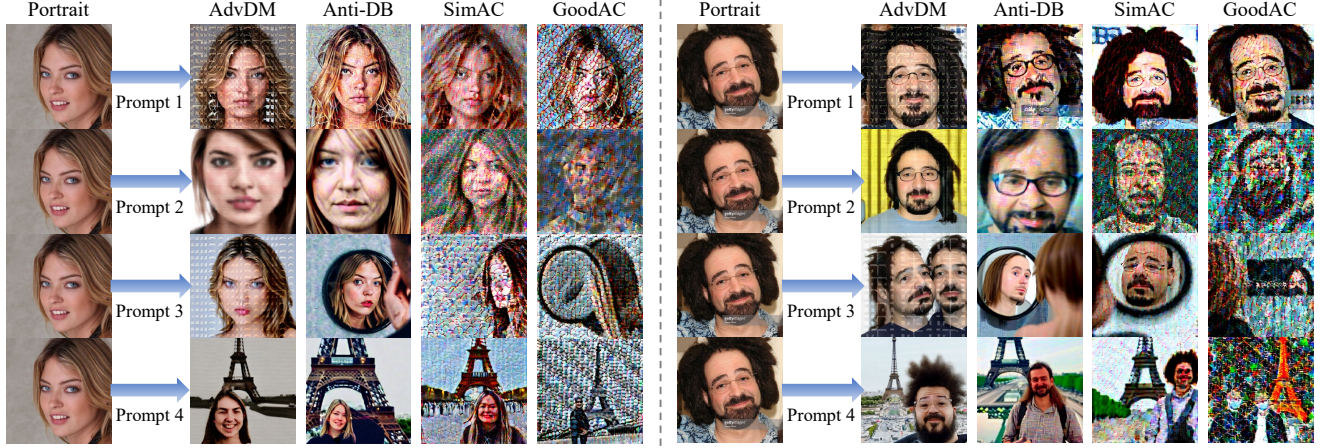


Figure 3. Visualization results (under four prompts) on CelebA-HQ and VGGFace2. Prompt 1-4 are respectively: “a photo of sks person”, “a dslr portrait of sks person”, “a photo of sks person looking at the mirror” and “a photo of sks person in front of Eiffel Tower”.

Table 3. CelebA-HQ Dataset Results: This table presents the performance comparison of different components under “a photo of sks person in front of eiffel tower”.

Method	“a photo of sks person in front of eiffel tower”			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
SimAC	0.04	71.87	43.99	0.12
$\mathcal{L}_{perceptual}$	0.01	83.33	41.80	0.02
$\mathcal{L}_{edge}$	0.09	76.66	43.86	0.13
$\mathcal{L}_{perceptual} + \beta\mathcal{L}_{edge}$	0.02	96.67	44.99	0.02

is also reduced by over 50%, and FDFR exceeds 90%.

(2) Our method demonstrates strong resistance to concept transfer. In the scenario involving concept transfer to the Eiffel Tower scenario, our method significantly outperforms even the best-performing method, SimAC, in both ISM and FDFR metrics. This indicates that GoodAC disrupts a set of spatially structured perceptual features, thereby impairing the interaction between the *sks person* concept and the scenario features during the concept transfer process, making such transfer more difficult.

(3) Our method also performs remarkably well in resisting semantic theft. As shown in Figure 3, our approach achieves superior anti-customization effects across different scenarios. Specifically, by precisely targeting and disrupting facial attributes, the customized images generated using our method contain almost no discernible facial details. In contrast, other methods still retain varying degrees of facial detail, further validating the effectiveness of our approach.

4.3. Ablation Studies

In this section, we conducted several ablation studies to further investigate the contributions of some crucial components and hyper-parameters in our method.

Table 4. CelebA-HQ Dataset Results: This table presents the performance comparison of different β values under “a photo of sks person in front of eiffel tower”.

β	“a photo of sks person in front of eiffel tower”			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
1.0	0.01	90.00	44.53	0.06
1.5	0.02	83.33	43.47	0.10
1.65	0.08	90.00	46.32	0.06
1.7	0.09	36.66	41.3	0.39
1.625	0.02	96.67	44.99	0.02

4.3.1. Ablation Study on Loss Function Components

To evaluate the effectiveness of the proposed loss components, we conduct experiments on the CelebA-HQ dataset and the Eiffel Tower scenario, comparing our method with SimAC, the strongest existing AC baseline. As shown in Table 3, a key observation is that incorporating either $\mathcal{L}_{perceptual}$ or $\mathcal{L}_{edge}$ individually already improves anti-customization performance, and their combination yields further gains. This validates our hierarchical design, where global and local strategies are collaboratively employed to achieve robust anti-customization. Furthermore, the individual improvements obtained by each loss function differ. Specifically, $\mathcal{L}_{perceptual}$, which targets global feature correlations, introduces large-scale perturbations by disrupting the spatial dependencies of perceptual features. When used alone, it already outperforms SimAC. In contrast, $\mathcal{L}_{edge}$ focuses on fine-grained and identity-specific facial attributes, acting as a guide that directs perturbations toward key semantic features. Although $\mathcal{L}_{edge}$ alone generates relatively subtle perturbations and performs slightly worse than SimAC, its combination with $\mathcal{L}_{perceptual}$ significantly enhances the overall anti-customization effectiveness.

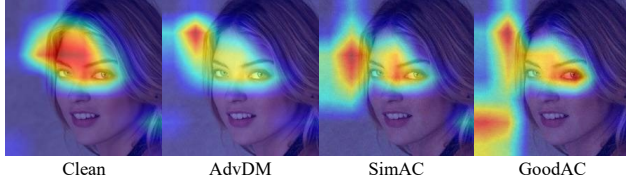


Figure 4. Attention maps before and after attack.

4.3.2. Ablation Study on Key Hyperparameter

In addition, we investigate the influence of the hyperparameter β , which balances the two loss terms, on anti-customization performance. On both the CelebA-HQ dataset and the Eiffel Tower scenario, we test β values of 1.0, 1.5, 1.65, 1.7, and 1.625. As shown in Table 4, the effectiveness of our GoodAC initially increases with larger β values, reaching a peak and then dropping rapidly beyond the optimal point. This behavior can be attributed to the fact that a smaller $\mathcal{L}_{edge}$ weight allows it to serve as a guiding component that enhances attack effectiveness. However, when $\mathcal{L}_{edge}$ becomes dominant, the perturbation focuses excessively on localized facial details while neglecting global feature disruption, thereby weakening the overall anti-customization impact. These results further support the necessity of maintaining a balance between global and local strategies for optimal collaboration. Based on this study, we identify $\beta = 1.625$ as the optimal setting.

5. Discussion

In this section, we further analyze the effectiveness of GoodAC by visualizing attention maps and feature maps, aiming to better understand how the global-local collaborative perturbation mechanism functions.

We first visualize the attention maps of Sample 5 from the CelebA-HQ dataset, including the clean image and the adversarial examples generated by AdvDM, SimAC, and GoodAC. The attention maps are generated using CAM [44] with the ResNet50 backbone, and the analysis is performed for the “woman” label. As shown in Figure 4, we observe that the attention maps of AdvDM and SimAC are more concentrated on specific regions, while the attention map of GoodAC is more dispersed. This observation indicates that global feature correlation disruption effectively impairs the spatial perception ability of customization models, resulting in scattered attention distribution that hinders key concept extraction from input images. Moreover, the degree of attention dispersion among different methods aligns with the defense effectiveness shown in Table 1 and Table 2, providing strong evidence for the effectiveness of the global feature correlation disruption strategy.

Additionally, we directly visualize the intermediate feature maps of the U-Net model to evaluate the effectiveness

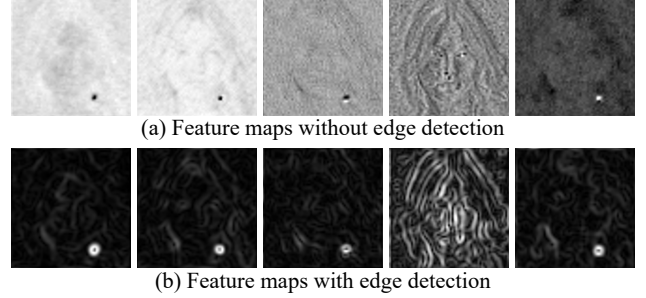


Figure 5. The feature maps w/o, w/ edge detection operator in boundary loss of sample 5 in CelebA-HQ dataset, best in view.

of the local facial attribute distortion strategy. Specifically, we select the feature maps from the last up-sampling block of the U-Net, using both the clean image of Sample 5 from the CelebA-HQ dataset and its corresponding adversarial image as input. As illustrated in Figure 5, it is evident that without the guidance of local facial attributes and edge detection, the perturbation tends to originate from the background and the entire facial area, resulting in a uniform distribution. In contrast, when local facial attribute guidance and edge detection are applied, the perturbation becomes concentrated on key facial attributes.

In summary, the above analyses demonstrate the targeted effectiveness of GoodAC on both global and local levels, which enables strong defense performance against customization methods.

6. Conclusion

In summary, we propose GoodAC, which achieves significant progress in enhancing the effectiveness of anti-customization (AC) methods by collaboratively disrupting global perceptual correlations and distorting local facial attributes. Our GoodAC demonstrates strong defense capability by resisting both concept transfer and semantic theft, outperforming existing state-of-the-art anti-customization approaches. Experimental results across various Latent Diffusion Models and customization methods show that our method provides superior protection against unauthorized facial usage, thereby advancing the standards of privacy and security in personalized image synthesis.

Despite these promising advancements, our approach still has some limitations. For instance, GoodAC has only been evaluated on open-source, fine-tuning-based models, and its performance on proprietary commercial models or non-fine-tuned models remains unexplored. Moreover, the GoodAC framework, like other existing anti-customization methods, might be vulnerable to specially designed customization techniques, pointing to a critical future research direction — the development of robust anti-customization (AC) strategies, which deserves further investigation.

References

- [1] Stability AI. Stable diffusion 2-1. <https://huggingface.co/stabilityai/stable-diffusion-2-1>, 2023.
- [2] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3389–3398, 2018.
- [3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.
- [4] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 32, 2019.
- [5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [6] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos P Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2021.
- [7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [8] Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems*, 33:8270–8283, 2020.
- [9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [14] Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. An empirical study of llama3 quantization: From llms to mllms. *Visual Intelligence*, 2(1): 36, 2024.
- [15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv: Machine Learning, arXiv: Machine Learning*, 2013.
- [17] Simin Li, Shuang Zhang, Guojun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, and Xianglong Liu. Towards benchmarking and assessing visual naturalness of physical world adversarial attacks. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12324–12333, 2023.
- [18] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 20763–20786, 2023.
- [19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.
- [20] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
- [21] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, page 4695–4708, 2012.
- [22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [23] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *CVPR Workshops*, 2017.
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- [27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

- [28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [29] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023.
- [30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [31] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [32] Samyakh Tukra, Frederick Hoffman, and Ken Chatfield. Improving visual representation learning through perceptual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14486–14495, 2023.
- [33] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023.
- [34] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [35] Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12047–12056, 2024.
- [36] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1722–1740, 2024.
- [37] Jiakai Wang, Xianglong Liu, Zixin Yin, Yuxuan Wang, Jun Guo, Haotong Qin, Qingtao Wu, and Aishan Liu. Generate transferable adversarial physical camouflages via triplet attention suppression. *International Journal of Computer Vision*, 132(11):5084–5100, 2024.
- [38] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [39] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.
- [40] Yanni Xue, Haojie Hao, Jiakai Wang, Qiang Sheng, Ren-shuai Tao, Yu Liang, Pu Feng, and Xianglong Liu. Vision-fused attack: Advancing aggressive and stealthy adversarial text against neural machine translation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6606–6614. International Joint Conferences on Artificial Intelligence Organization, 2024. Main Track.
- [41] Zixin Yin, Jiakai Wang, Yisong Xiao, Hanqing Zhao, Tianlin Li, Wenbo Zhou, Aishan Liu, and Xianglong Liu. Improving deepfake detection generalization by invariant risk minimization. *IEEE Transactions on Multimedia*, 26:6785–6798, 2024.
- [42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [43] Xingyu Zheng, Xianglong Liu, Haotong Qin, Xudong Ma, Mingyuan Zhang, Haojie Hao, Jiakai Wang, Zixiang Zhao, Jinyang Guo, and Michele Magno. Binarydm: Accurate weight binarization for efficient diffusion models, 2024.
- [44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.