Harnessing Global-Local Collaborative Adversarial Perturbation for Anti-Customization

Supplementary Material

A. Structure of Supplementary Material

This supplementary material is divided into two parts to provide a more comprehensive understanding of the details of our proposed method and its experimental results. We start by providing an overview of our GoodAC in Section B, accompanied by pseudocode for clarity. In Section C, we added some experimental details and results. We first explained the calculation details of the evaluation metrics, then we further illustrate the effectiveness of our GoodAC across various prompts, and finally evaluate the effectiveness of our method across different noise budgets, model version discrepancies, and different customization methods.

B. The Overview of GoodAC

The GoodAC method, as shown in Algorithm 1, aims to generate a corresponding perturbation δ for each clean image x in the dataset $\mathcal{X}_\mathcal{B}$ to be protected. A iterative strategy is employed to perturb the images in $\mathcal{X}_\mathcal{B}$. In each iteration, we first fine-tune the model using the clean reference dataset $\mathcal{X}_\mathcal{A}$ to enable the model to learn the correct concepts of the images. Then, we compute three types of losses to obtain the adversarial perturbation δ, as follows: ❶ $\mathcal{L}_{cond}$: Compute and maximize the conditional loss $\mathcal{L}cond$ of the current image under the original LDM so that the perturbed image deviates from the correct concept. ❷ $\mathcal{L}_{GoodAC}$: As described in the main paper, block angle transformations are first applied to disrupt the spatial correlations of perceptual features, followed by edge detection to extract fine facial attributes for targeted distortion. The loss function $\mathcal{L}_{GoodAC}$ is calculated based on these operations. ❸ $\mathcal{L}_{feat}$: Given the feature-level attack advantage of SimAC, we compute $\mathcal{L}_{feat}$ and incorporate it into the total loss, with its weight adjusted by the parameter γ to prevent perturbations from affecting non-critical facial attributes.

Finally, we use these three loss functions to generate the final adversarial perturbation δ, constrain its L_∞ norm within η, and add it to the image to produce the final perturbed image. After the iterative process is completed, all images in $\mathcal{X}_\mathcal{B}$ are transformed into their perturbed versions, resulting in the protected dataset ${\mathcal{X}_\mathcal{B}}^{(adv)}$ and the corresponding perturbations $\delta^{(adv)}$.

Algorithm 1 GoodAC algorithm

Require:
$\mathcal{X}_\mathcal{A}$: reference clean dataset
$\mathcal{X}_\mathcal{B}$: clean dataset to be protected
θ: parameters of the customized LDM
β and γ: weight factors
κ and ν: number of blocks and maximum rotation angle
η: adversarial perturbation budget
Ensure:
${\mathcal{X}_\mathcal{B}}^{(adv)}$: protected image dataset
$\delta^{(adv)}$: adversarial perturbation of each image in ${\mathcal{X}_\mathcal{B}}^{(adv)}$
Begin:
$\theta' \leftarrow \theta$, ${\mathcal{X}_\mathcal{B}}^{(adv)} \leftarrow \mathcal{X}_\mathcal{B}$, $\delta^{(adv)} \leftarrow \{\mathbf{0}, ..., \mathbf{0}\}$
for each training step t in $\{1, \ldots, 50\}$ **do**
  $\theta' \leftarrow \arg\min_{\theta'} \sum_{x' \in \mathcal{X}_A} \mathcal{L}_{db}(\theta', x')$
  for each image x in ${\mathcal{X}_\mathcal{B}}^{(adv)}$ **do**
    calculate $\mathcal{L}_{cond}(\theta', x)$
    calculate $\mathcal{L}_{GoodAC}(\theta', x, \beta, \kappa, \nu)$
    calculate $\mathcal{L}_{feat}(\theta', x)$
    $\mathcal{L}_{all} \leftarrow \mathcal{L}_{cond} + \mathcal{L}_{GoodAC} + \gamma \cdot \mathcal{L}_{feat}$
    for each PGD step p in $\{1, \ldots, 6\}$ **do**
      $\delta \leftarrow \alpha \cdot \text{sign}(\nabla_{x_{adv}} \mathcal{L}_{all})$
    end for
    clamp $\|\delta\|_\infty$ to η
    $x \leftarrow x + \delta$
    $\delta^{(adv)}[x.index] \leftarrow \delta$
  end for
  $\theta' \leftarrow \arg\min_{\theta'} \sum_{x' \in \mathcal{X}_B} \mathcal{L}_{db}(\theta', x')$
end for
return ${\mathcal{X}_\mathcal{B}}^{(adv)}$, $\delta^{(adv)}$
End

C. Experiment

In this section, we added some experimental details and provide additional experiments to comprehensively evaluate our GoodAC. We first add the calculation details of the evaluation metrics in subsection C.1. Then, in subsection C.2, we employ a wider range of prompts to demonstrate the effectiveness of GoodAC in resisting concept transfer in subsection C.2. Subsequently, we analyze the impact of different noise budgets on the attack performance in subsection C.3. Finally, we evaluate the robustness of our GoodAC against discrepancies between model versions in subsection C.4, and evaluate the ability of our GoodAC across various customization methods in subsection C.5.

C.1. Calculation Details of Evaluation Metrics

The Face Detection Failure Rate (FDFR) measures the proportion of generated images from successfully disrupted DreamBooth models that contain no detectable face. We

Figure 6. Visualization results across various prompts on CelebA-HQ dataset. Our GoodAC maintains excellent resistance to concept transfer in most scenarios.

use the RetinaFace detector [5] for face detection. For images where a face is detected, we calculate the Identity Score Matching (ISM) by extracting face recognition embeddings using the ArcFace recognizer [6], and computing the cosine distance between the embedding of the generated image and the average embedding of the user's clean image set. This metric reflects the identity similarity. Additionally, we use SER-FQA [21, 31], a recently proposed image quality assessment metric specifically tailored for facial images, and BRISQUE, a classical and widely adopted no-reference image quality metric. Since ArcFace embeddings may not fully capture differences in visual content or facial attributes, we further extract general visual features using a ResNet50 encoder [11] to enhance similarity measurement.

C.2. More Results on Different Prompts

To further evaluate the effectiveness of our proposed GoodAC method in resisting concept transfer across various prompts, we conduct additional experiments using a more diverse set of prompts with finer-grained scenarios. These prompts are as follows:

- Prompt A: *"a photo of sks person reading a book"*
- Prompt B: *"a photo of sks person in a kitchen cooking"*
- Prompt C: *"a photo of sks person at a business meeting"*
- Prompt D: *"a photo of sks at a train station"*
- Prompt E: *"a photo of sks person in a forest hiking"*
- Prompt F: *"a photo of sks person wearing a Nike shirt"*
- Prompt G: *"a photo of sks person hiking in mountains"*
- Prompt H: *"a photo of sks person at a yoga class"*

The results are shown in Figure 6. It can be seen that in different scenarios, the new images generated by the model based on the protected images are covered with heavy textures and fail to capture precise facial attributes. This effectively demonstrates the efficacy of our proposed perceptual feature correlation disruption strategy and precise facial attribute distortion strategy. These techniques ensure that the protected images generated by GoodAC maintain excellent resistance to concept transfer in most scenarios.

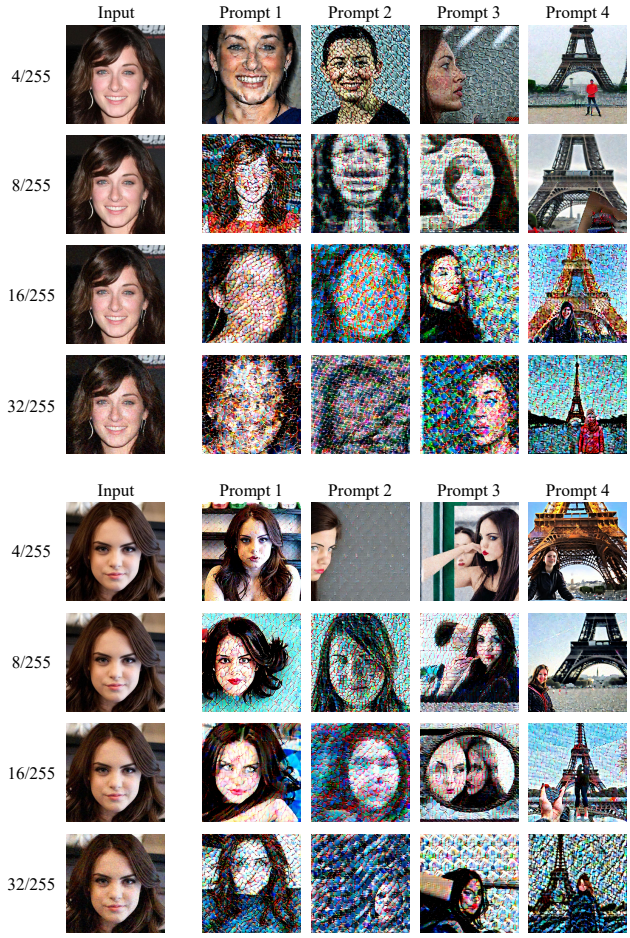

Figure 7. Results of different noise budgets under original four prompts on CelebA-HQ dataset.

C.3. Extra Analysis on Different Noise Budgets

It is crucial to understand the impact of different noise budgets on the performance of our method. In the main paper, we set a standard noise budget of 16/255 for GoodAC and other comparison methods, as excessive perturbation on the original image can severely degrade its quality. Here, we conduct further analytical experiments using different noise constraints, including 4/255, 8/255 and 32/255. We follow the settings in the main paper and conduct experiments using the following four prompts:

- Prompt 1: *"a photo of sks person"*
- Prompt 2: *"a dslr portrait of sks person"*
- Prompt 3: *"a photo of sks person looking at the mirror"*
- Prompt 4: *"a photo of sks person in front of eiffel tower"*

The results are shown in Figure 7. With an increasing perturbation budget, the textures overlaying the generated images become more chaotic, and facial attributes become less distinct. Thus, it can be concluded that the anti-customization effect improves as the noise budget increases.

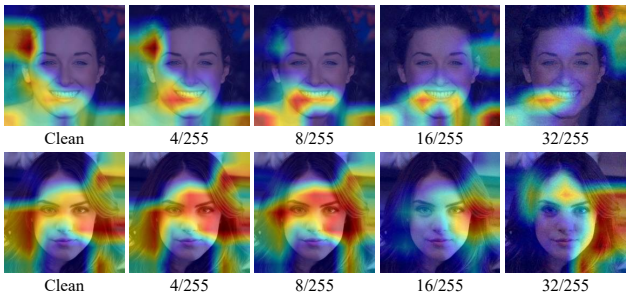

Figure 8. Different noise budgets attention maps before and after anti-customization of our GoodAC.

Additionally, we visualized the attention maps of anti-customized examples under different noise budgets to observe the impact of noise budgets on global feature correlation disruption. As shown in Figure 8, we can see that at a low budget, the model's attention, similar to the clean image, focuses on facial features such as the mouth and eyes. However, when the budget increases to 16/255, the attention shifts to the image background and other irrelevant areas, thereby increasing the disruption of perceptual feature correlations. From this, we can conclude that as the budget increases, the degree of disruption to perceptual feature correlations rises, leading to a more effective anti-customization.

Furthermore, we provide quantitative data to demonstrate the impact of noise budgets on anti-customization effectiveness. In our experiments, we used varying noise budgets based on the GoodAC method on the CelebA-HQ dataset to assess anti-customization effectiveness. The evaluation is based on four key metrics: ISM (Identity Similarity Metric), FDFR (Face Detection Failure Rate), BRISQUE, and SER-FQA (Semantic and Entity-based Realism Fidelity Quality Assessment). We observe that as the noise budget increases, ISM and SER-FQA decrease, indicating higher dissimilarity between the generated image and the original, as well as reduced perceived realism. These trends align with the goals of anti-customization, where lower ISM values are preferable, as they signify less similarity between the generated image and the original face, and lower SER-FQA values are desirable, as they suggest lower likelihood of appearing realistic or recognizable by facial recognition systems. In contrast, FDFR and BRISQUE exhibit the opposite trend. Specifically, higher FDFR values indicate that generated images are more likely to evade face detection models, beneficial for anti-customization. Overall, the results are shown in Table 1, it can be seen that increasing the noise budget leads to greater degradation of the generated image quality.

Table 5. Different noise budgets based on GoodAC on CelebA-HQ dataset, where lower ISM and SER-FQA are better and higher FDFR and BRISQUE are better.

budget	*"a photo of sks person"*			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
4/255	0.11	56.67	38.66	0.31
8/255	0.06	83.33	38.70	0.12
16/255	0.03	96.67	39.75	0.02
32/255	0.01	99.99	40.33	0.01
budget	*"a dslr portrait of sks person"*			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
4/255	0.05	66.67	25.94	0.23
8/255	0.03	93.33	39.88	0.04
16/255	0.01	99.99	43.21	0.01
32/255	0.01	99.99	45.55	0.01
budget	*"a photo of sks person looking at the mirror"*			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
4/255	0.07	46.67	37.48	0.35
8/255	0.06	73.33	42.47	0.17
16/255	0.02	96.67	44.11	0.01
32/255	0.01	99.99	44.90	0.01
budget	*"a photo of sks person in front of eiffel tower"*			
	ISM↓	FDFR↑	BRISQUE↑	SER-FQA↓
4/255	0.11	30.01	38.11	0.43
8/255	0.11	65.75	40.23	0.38
16/255	0.02	96.67	44.99	0.02
32/255	0.01	99.99	45.81	0.01

C.4. Extra Analysis on Model Mismatch

To evaluate the sensitivity of our GoodAC to discrepancies between model versions, which we refer to as model mismatch, we tested its anti-customization performance across different iterations of the Stable Diffusion model. We follow the settings in the main paper and conduct experiments using the original four prompts as above.

Specifically, we trained our method using Stable Diffusion v2.1 and evaluated it on both Stable Diffusion v1.4 and Stable Diffusion v2.1. As shown in Figure 9, when tested on SDv1.4, the facial features in the generated images under prompts 2, 3, and 4 are largely lost, indicating that our method effectively resists concept migration across different model versions. When using prompt 1, we observe that the faces still exhibit distortion, and the protected facial features are not fully replicated. Therefore, we can conclude that our method maintains strong anti-customization performance even under model mismatch conditions.

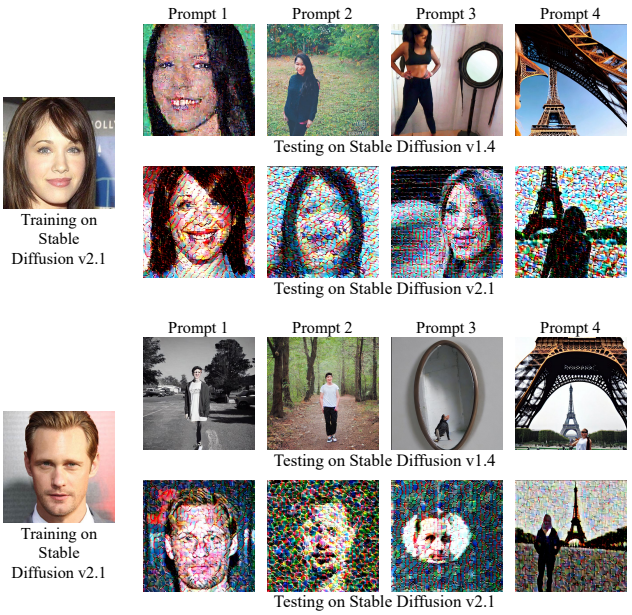

Figure 9. Results of models mismatch under original four prompts on CelebA-HQ dataset.

Figure 10. Results of customization mismatch under four prompts on CelebA-HQ dataset with SimAC and GoodAC. The training is based on Dreambooth and the customization test is based on Lora. The experiment simulates an attacker using an unknown customization method.

C.5. Extra Analysis on Customization Method Mismatch

To simulate scenarios where attackers employ unknown or different customization techniques, we trained our GoodAC method based on Dreambooth and tested it against models customized using LoRA. The results, as shown in Figure 10, indicate that when attackers use LoRA, SimAC performs poorly across all prompts, largely losing its anti-customization capability and allowing facial features to be almost fully restored. The reason for this is that SimAC overly relies on the model's gradient information and fails to account for the limitations of image features, leading to a strong coupling between the adversarial perturbation and the model. In contrast, GoodAC maintains efficient anti-customization performance, effectively distorting facial features across all prompts. Thus, even when there is a mismatch between the customization methods used during training and testing, GoodAC effectively resists customization. Therefore, we conclude that our method provides robust anti-customization performance against different customization strategies employed by attackers.