Paper Summary:

Enhancing Facial Privacy Protection via Weakening Diffusion Purification

Afreen Alam

CS7064: Trustworthy Machine Learning

January 9, 2026

Summary of the Paper

To combat the misuse of Automatic Face Recognition (AFR) systems to identify, track and monitor individuals without their consent, the authors of this paper propose a novel method for facial privacy protection by generating protected images which are visually similar to the original image but can fool AFR systems by impersonating a different target identity.

The key problem addressed by the authors in this paper is that existing diffusion-based facial privacy methods, like DiffProtect, suffer from diffusion purification effect, where the protective adversarial perturbations are unintentionally removed during the denoising steps, resulting in poor Protection Success Rate (PSR). Additionally, the authors note that the modification of the semantic code causes the protected face to structurally resemble the target face, leading to visually noticeable alterations from the original image and thus resulting in poor visual output image quality.

To overcome these challenges, the authors propose a framework which involves a two-stage learning strategy within a Latent Diffusion Model (LDM). In the first stage, unconditional embeddings are learned per-timestamp. These learned unconditional embeddings are used as null-text guidance in the reverse diffusion process which not only allow retention of fine textures and structural details but more importantly weaken the diffusion purification effect by enabling identity-related adversarial modifications to persist. This step effectively improves both the visual quality and protection capability of the generated image. The second stage integrates self-attention guidance to maintain the structural consistency between the original and generated image. By aligning the self-attention maps before and after adding the perturbation, the model is better able to preserve the geometric shape and structural details of the original image and therefore maintain high visual quality.

Extensive experiments using two datasets and four black-box FR models were carried out and compared with four noise-based, four makeup-based, and one diffusion-based facial privacy protection techniques. The results demonstrate that the proposed method outperforms existing state-of-the-art methods by achieving higher PSRs and maintaining higher image quality. It achieves a substantial improvement in average PSR over noise-based and makeup-based methods, 30% and 1.5%, respectively. Qualitative studies also show that the proposed method generates more natural looking images compared to the other methods.

In conclusion, the method proposed by the authors of this paper is able to effectively generate high quality adversarial images which are imperceptible to human eyes but can successfully deceive Automatic Facial Recognition systems.

Strengths and Weaknesses

Strengths

1. Learning unconditional embeddings per-timestamp for null-text guidance to weaken the diffusion purification effect is a novel approach to retain adversarial perturbations during the denoising steps.

2. Integration of self-attention maps to maintain structural consistency of the original and generated image is a well-reasoned way to ensure high visual quality and directly addresses the trade-off between semantic code modification strength and protection capability which existing diffusion-based methods face.

3. The proposed method is independent of any reference image or text-based prompts unlike existing methods which makes it more practical for real-world scenarios.

4. The authors addressed ethical concerns of using real images as target images and instead chose to use synthesized target images for impersonation to avoid misusing a real person's similarity.

Weaknesses

1. The authors did not define some of the evaluation metrics such as FID and PSNR. Although references were given, a little explanation of the metrics would have made understanding the evaluation criteria better as well as interpreting the results in the Tables.

2. The authors claim that their proposed method achieves better PSNR performance compared to all the other methods. However, Table 2 shows that that is not the case. TIP-IM and Adv-makeup have higher PSNR scores. So, this is an overstatement on the authors' part. Although, as a reader I could infer that their method achieves a better trade-off between PSNR and PSR, it would have been better if the authors made that clearer.

Questions

1. Since only two datasets are used, is there a chance that the protection effectiveness or the visual quality might not be the same across various demographics?

2. Do the target images used have to be carefully selected ensuring that the visual similarity between the original image and the target image vary quite significantly? What happens if the target and original image are more similar in terms of performance impact?

3. The method requires per-timestep optimization of null-text embeddings and iterative adversarial optimization in latent space. How does this scale with image resolution and dataset size in terms of complexity?

4. In the Supplementary materials, the authors mentioned their plan to replace the current surrogate model-based training paradigm. How would that work?

5. To improve transferability the method uses multiple surrogate models, but how robust would it be to unseen or proprietary or newer FR systems?