Adversarial Privacy-preserving Filter

# Introduction

While widely adopted in practical applications, face recognition has been critically discussed regarding the malicious use of face images and the potential privacy problems, e.g., deceiving payment system and causing personal sabotage. Online photo sharing services unintentionally act as the main repository for malicious crawler and face recognition applications. This work aims to develop a privacy-preserving solution, called Adversarial Privacy-preserving Filter (APF), to protect the online shared face images from being maliciously used.We propose an end-cloud collaborated adversarial attack solution to satisfy requirements of privacy, utility and nonaccessibility. Specifically, the solutions consist of three modules: (1) image-specific gradient generation, to extract image-specific gradient in the user end with a compressed probe model; (2) adversarial gradient transfer, to fine-tune the image-specific gradient in the server cloud; and (3) universal adversarial perturbation enhancement, to append image-independent perturbation to derive the final adversarial noise. Extensive experiments on three datasets validate the effectiveness and efficiency of the proposed solution. A prototype application is also released for further evaluation. We hope the end-cloud collaborated attack framework could shed light on addressing the issue of online multimedia sharing privacy-preserving issues from user side.

The contributions of the paper can be summarized as follows:

1. We designed an adversarial privacy-preserving filter to preserve users' portrait privacy from malicious face recognition crackers without affecting their photo sharing experience. Adversarial attack naturally meets the two basic requirements of privacy and utility.
2. We proposed an end-cloud collaborated adversarial attack framework, which addresses the additional non-accessibility requirement to guarantee the original image only accessible to users' own device end. The compatible performance of this two-stage attack with the traditional one-stage attack also provides a novel perspective to understand the adversarial attack problem.
3. We integrated the universal adversarial perturbation with image-dependent perturbation to obtain improved privacy-preserving capability. This sheds light on alternative way to exploit adversarial examples.
4. We conducted extensive experiments to validate the effectiveness and efficiency of the proposed solution framework. A prototype of adversarial privacy-preserving filter is further carried out and released for evaluation.

# Framework

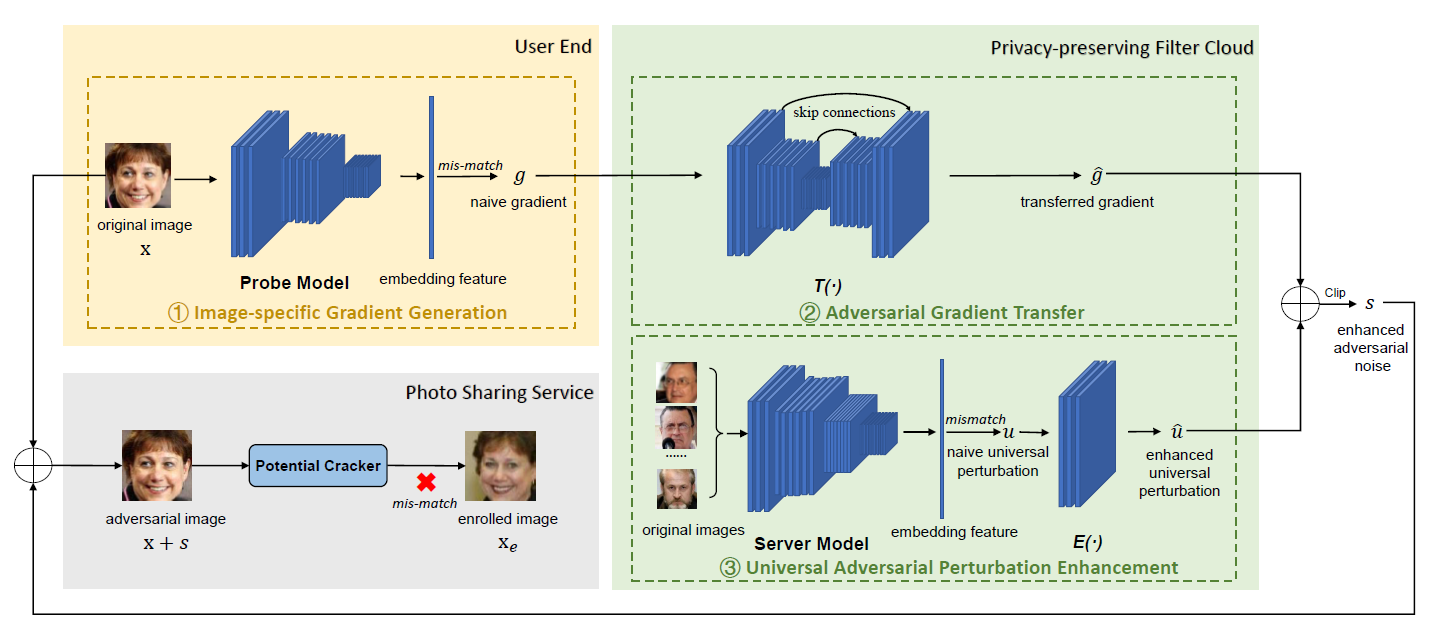


Fig.1 The proposed adversarial privacy-preserving filter framework.

# Paper

https://dl.acm.org/doi/abs/10.1145/3394171.3413906