PROTECTING IMAGES FROM MACHINE LEARNING MODELS USING NATURALLY CAMOUFLAGED ADVERSARIAL PATCHES.

Huynh Chi Nhan

University of Information Technology - National University of Vietnam

Nguyen Ho Nam

Department of Computer Science

What ?

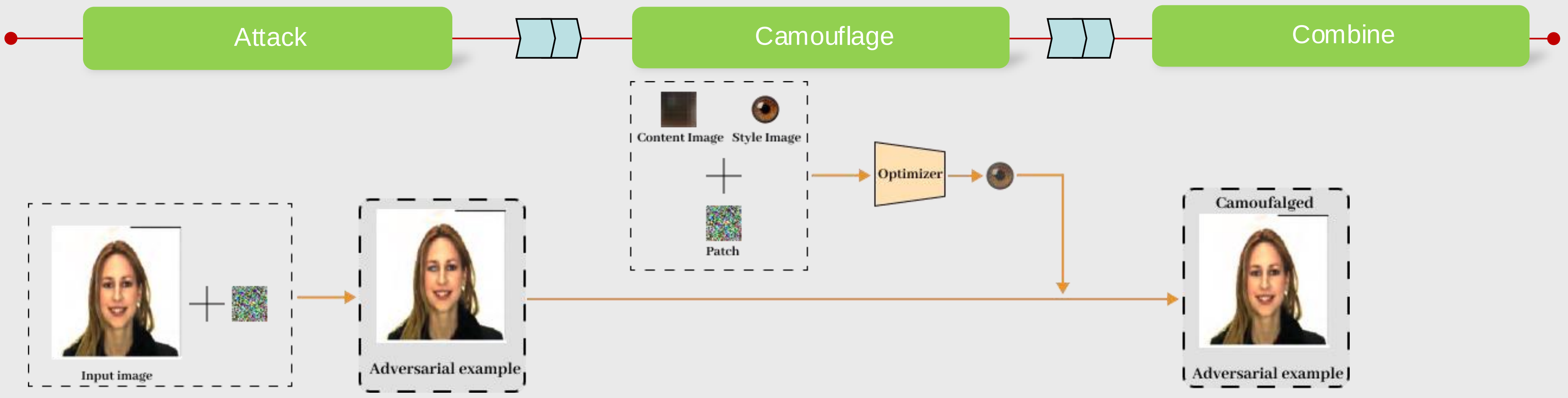
The study on generating Camouflaged Adversarial Patches aims to achieve the following objectives:

- Experiment with the implementation of Adversarial Patches to evaluate their performance on the YOLOv8 detection model.
- Enhance the content and style of the patches using Style Transfer techniques.
- Provide foundational knowledge and mathematical principles related to Adversarial Patches and Style Transfer.

Why ?

- Personal images on the internet can easily be exploited to create Deepfake videos. So, how can users protect their personal images?
- To address this issue, we propose using Camouflaged Adversarial Patches, a method that disrupts the facial extraction process of users. This solution not only safeguards personal images but also preserves the aesthetic and natural appearance of the original photos.

Overview



Description

1. Dataset and detection model

For classification, we experiment on the YOLOv8 model on the CelebA dataset.

Figure 1. CelebA dataset.

2. Adversarial Patches

Since we do not have access to the parameters of the YOLO model, we construct the **Adversarial Patch** based on the model's input and output. The eyes are a critical region in face detection; therefore, we create a **Mask** based on the size and position of the eyes. To cause the model to misclassify the ground truth of the bounding box and class, we optimize a patch (**Patch**) to maximize the loss of the object detector with respect to the ground truth label and bounding box, when the patch is applied through a predefined function.

$$\hat{P}_u = \arg\max_{P} \mathbb{E}_{x,s}\left[L(A(x,s,P);\hat{y},\hat{B})\right]$$

During back-propagation, we update the pixels of **Patch**.

3. Style Transfer

We provide both a style image and a content image. The style image is sourced externally, while the content image represents the region of the original image covered by the patch.

In a neural network, style optimization is performed using features extracted from the lower-dimensional layers, whereas content optimization leverages high-dimensional features from deeper layers.

The optimization process uses cross-entropy as the primary loss function and focuses on two key components:

1.Content Loss: Ensures that the structural information of the original image (or the patch area) remains recognizable.

$$L_{\text{content}}(C,\delta) = \frac{1}{2}\sum_{i,j}\left(C_{ij}^{l} - \delta_{ij}^{l}\right)^2$$

2. Style Loss: Transfers the texture and visual features of the selected style image (e.g., rust texture) onto the patch. Style features are extracted from the style image using the Gram matrix of the convolutional layers in a pre-trained neural network.

$$L_{\text{style}}(S,\delta) = \sum_{l}\frac{1}{N_l M_l}\sum_{i,j}\left(G_{ij}^{l}(\delta) - G_{ij}^{l}(S)\right)^2$$

4. Camouflaged Adversarial Patch

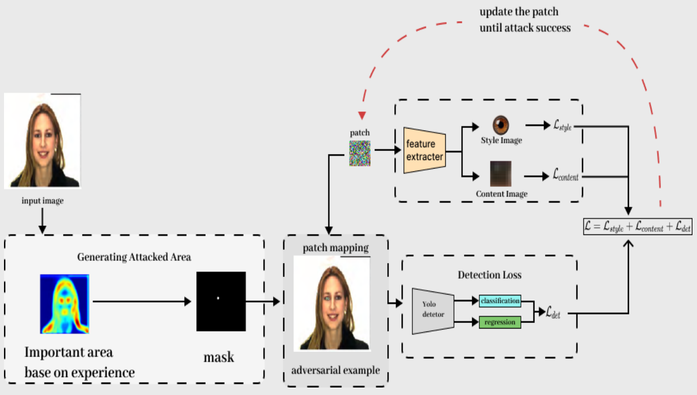


Figure 2. Diagram of Camouflaged Adversarial Patch.

NII

Huynh Chi Nhan – University of Infomation Technology

TEL : 0376856852 Email : 22520996@gm.uit.edu.vn