ABSTRACT
This work tackles the Pixel Privacy task put forth by MediaEval 2019. Our goal is to decrease the accuracy of a classification algorithm while preserving the original image quality. We use the fast gradient sign method, which normally has a corrupting influence on image appeal, and devise two methods to minimize the damage. The first approach uses a map that is a combination of salient and flat areas. Perturbations are more noticeable in these locations, and so are directed away from them. The second approach adds the gradient of an aesthetic algorithm to the gradient of the attacking algorithm to guide the perturbations towards a direction that preserves appeal. We make our code available at: https://git.io/JesXr.

1 INTRODUCTION
The Pixel Privacy task, introduced by MediaEval [11], aims at developing methods for manipulating images in a way that fools automatic scene classifiers. As an added constraint, the images should not exhibit a decrease in aesthetic quality; it is even encouraged that their appeal increases. The organizers made available the Places365-Standard data set [10] along with a pre-trained scene threat model following the ResNet [3] for the task. In this work, the classification model will be referred to as “threat model.”

The contribution of image enhancement techniques in privacy protection has been previously explored [1], showing that even popular filters used in social media have a cloaking effect against geo-location algorithms. A more recent work by Liu et al. [6] proposed a perturbation-based approach (white-box) and a transfer style approach (black-box). Similar to the first module in that work, we propose two perturbation-based approaches and explore ways to localize the perturbations in a manner that does not reduce appeal.

2 APPROACH
We developed two approaches, both of which utilize the FGSM adversarial attack [2]. This method uses the gradient of the threat model and changes the pixel values by nudging them towards the direction that maximizes the loss. The magnitude of this perturbation is represented by the value $\epsilon$, which is essentially the strength of the attack. Note that we used the threat model’s gradients, making this a white-box setting.

To mitigate the corrupting effect the FGSM attack has on the image, our first approach uses a combination of saliency maps and Sobel filters. The goal is to localize the perturbations in areas that are less obvious to observers. Our second approach constrains the attack towards a direction that is more aesthetically pleasing.

2.1 Salient Defence
We combine two maps: one is a measure of saliency and the other a measure of flatness. Saliency areas are the ones that are more likely to attract the eye of an observer. We use a deep neural network that predicts saliency from images to detect these areas. In particular, we use a SalBCE [5] trained on the SALICON dataset [4]. Furthermore, perturbations become more obvious when they are located in flat areas. For this reason, we also used a Sobel filter [8], which detects areas where edges are more prevalent and Gaussian blurring ($\sigma=10$) to spread the detected edges, forming a map. The saliency map is reversed so that the pixels corresponding to salient areas are zeroed out. Then, pixels where the mean value is below average on the Sobel map (hence more likely to be a flat area) are also zeroed out. The resulting map is then multiplied by $\epsilon$ as well as the gradient of the aesthetic assessment algorithm (NIMA).

Figure 1: Maps used to constrain perturbations on less obvious areas. The reverse saliency map along with the Sobel map produce the final map.

1. Original Image
2. Sobel Map
3. Reverse Saliency Map
4. Final Map
2.2 Coupled Optimization
The second approach exploits the gradients of both the threat model and the aesthetics evaluation algorithm. The aesthetics evaluation in our case is the NIMA algorithm [9]. Since the networks differ significantly, the gradients need to be normalized to unit length in order to be brought to the same scale. Afterwards, NIMA’s gradient is subtracted from ResNet’s and as a result we get the sign of the total gradient and multiply that by $\epsilon$.

3 RESULTS AND ANALYSIS
In our initial experiments, we used a variety of $\epsilon$ values to investigate how it affects the accuracy of the threat model (Figure 2). Compared with the default FGSM attack, our approach is constrained on non-flat, non-salient areas, which naturally reduces the strength of the algorithm as less pixels are perturbed. We also note that additional manipulations on the image (tilt-shift filter in this case) further reduce the efficacy of those perturbations. The coupled optimization approach, on the other hand, has a higher impact on the accuracy of the threat model, as it manipulates all the pixels of the image. Note that we obtained these results using the original full size of each image from Places365.

On the other hand, images evaluated by the MediaEval team (Table 1) were first downsampled to $256 \times 256$ and the algorithms were applied afterwards. Note that this set includes only images that ResNet predicts successfully, therefore the initial accuracy ($\epsilon = 0$) is 100%. In that case it seems that the tilt-shift effect actually adds to the efficacy of the perturbations, bringing the accuracy of the threat model down.

To test NIMA’s sensitivity to perturbations, we attempted a vanilla FGSM attack with a very high $\epsilon = 0.15$ on a small subset (100) of the validation images. This type of attack effectively ruins the visual appeal; however, the NIMA score drops by only

\[
\begin{array}{cccc}
\text{Methods} & \epsilon & \text{Top-1 Acc.} & \text{NIMA Score} \\
\hline
\text{S. Defence} & 0.01 & 0.937 & 4.63 \\
& 0.05 & 0.735 & 4.58 \\
\text{S. Defence & tilt-shift} & 0.01 & 0.868 & 4.75 \\
& 0.05 & 0.458 & 4.54 \\
\text{C. Optimization} & 0.01 & 0.917 & 4.63 \\
& 0.05 & 0.458 & 4.54 \\
\text{Original Test Set} & - & 1.0 & 4.64 \\
\end{array}
\]

Table 1: Results on MediaEval test set. Top-1 accuracy refers to the prediction accuracy of the threat model (ResNet50 trained on Places365-standard data set). The NIMA Score column represents the average of the aesthetics scores.

4 DISCUSSION AND OUTLOOK
Our qualitative analysis showed that these perturbations, even when hidden, become more obvious on a high-resolution image. This stems from a rather obvious shortcoming of our salient defence algorithm: that saliency is subject to change after manipulations to the image. We believe that future effort should be targeted towards reducing the saliency of these perturbations. Furthermore, our coupled optimization approach effectively fools NIMA without truly preserving appeal. We believe that aesthetic algorithms that take into consideration low-level cues such as perturbations could improve the efficacy of this approach.
ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/15/SIRG/3283 and SFI/12/RC/2289.

REFERENCES


