Adversarial Image Perturbation for Privacy Protection
A Game Theory Perspective

Seong Joon Oh Mario Fritz Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
{joon,mfritz,schiele}@mpi-inf.mpg.de

Abstract

Users like sharing personal photos with others through social media. At the same time, they might want to make automatic identification in such photos difficult or even impossible. Classic obfuscation methods such as blurring are not only unpleasant but also not as effective as one would expect [28, 37, 18]. Recent studies on adversarial image perturbations *(AIP) suggest that it is possible to confuse recognition systems effectively without unpleasant artifacts. However, in the presence of counter measures against AIPs [7], it is unclear how effective AIP would be in particular when the choice of counter measure is unknown. Game theory provides tools for studying the interaction between agents with uncertainties in the strategies. We introduce a general game theoretical framework for the user-recogniser dynamics, and present a case study that involves current state of the art AIP and person recognition techniques. We derive the optimal strategy for the user that assures an upper bound on the recognition rate independent of the recogniser's counter measure. Code is available at* https://goo.gl/hgvbNK.

1. Introduction

People nowadays share massive amounts of personal photos through social media. Personal photos contain rich private information, *e.g.* about family members, travel destinations, and political activities. Together with recent developments in computer vision techniques [4, 11, 8, 27, 34], this results in increasing concerns that malicious entities employing computer vision technologies could extract private information from visual data.

Classical obfuscation techniques, such as face blurring and pixellisation, is not only unpleasant but also ineffective against convnet-based recognisers [28, 37, 18].

There have been recent studies on *adversarial image perturbations* (AIP): carefully crafted additive perturbations on the image that confuses a convnet while being nearly invisible to human eyes [36, 6, 21, 20]. AIPs are indeed promising as obfuscation techniques.

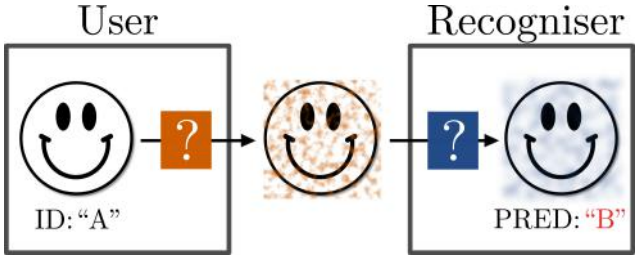

Figure 1: A game between a social media user and a recogniser over a photo. The user perturbs the image using orange strategy, trying to confuse the recogniser. The recogniser chooses blue strategy as a counter measure. They do not know which strategy is picked by the other.

However, it remains a question whether AIPs are still effective when counter measures are taken. For example, [7] proposed simple image processing tactics to counter the AIP effects (*e.g.* blurring by small amount). If furthermore the particular choice of counter measure is unknown, the best strategy is not obvious for the user.

Game theory provides useful tools for analysis when there exist uncertainties in the strategies for each player. We present a game theoretical framework to describe a system in which the user and recogniser strive for antagonistic goals: dis-/enabling recognition. This framework makes it possible to derive guarantees on the user's level of privacy, independent of the recogniser's counter measure, from an explicitly formulated set of assumptions. We include a case study of a person identification game, deriving the user's privacy guarantee with respect to the current state of the art AIP and person recognition methods.

This paper showcases the utility of game theory in understanding the user-recogniser dynamics. The framework can be extended beyond the particular settings considered. We believe this framework will further aid user-recogniser analyses in more diverse tasks and setups.

We list our contributions as follows:

- A game theoretic framework for studying the user-recogniser dynamics.
- Application of *adversarial image perturbation* (AIP) as an effective and aesthetic technique for person obfuscation.
- Novel robust and recogniser-selective AIPs.
- An empirical case study of the game theoretic framework, leading to the privacy guarantees for the user.

2. Related Work

Privacy and computer vision. While there exists a bulk of research on user privacy traditionally led by the security community [22, 39, 23, 19], studies on private content in visual data began only recently [37, 28, 18].

Wilber *et al*. [37] studied the performance of a commercial *face detector* under multiple face obfuscation methods (blur, darkening, camouflage glasses, etc.). Oh *et al*. [28] and McPherson *et al*. [18] studied the *face recognition* performance. In particular, [28] showed that current recognisers can adapt to obfuscation patterns. Above works conclude that recognisers can be robust against simple obfuscation methods like face blurring. In this work, we study a stronger obfuscation type: adversarial image perturbations.

Adversarial image perturbation (AIP). Szegedy *et al*. [36] first studied the phenomenon of adversarial instability of convnets: it is possible to generate invisible additive perturbations that completely fool a recogniser. The initial crafting algorithm was based on the L-BFGS [36]; more efficient first-order algorithms have been proposed [6, 31, 21, 12]. We review existing AIP algorithms and our novel variants conceptually and empirically.

Robust classification against AIPs. Some pre-convnet works considered enhancing general robustness of classifiers by training on adversarial data. Lanckriet *et al*. [13] trained a linear classifier on adversarial data constrained to a fixed mean and covariance for each class. Brückner *et al*. [3] introduced game theoretic concepts to formalise the adversarial training procedure. However, they limited their attention to simpler models: linear [13] or convex [3]. This work builds on a game theoretic framework which accommodates state of the art convnet models.

Since the advent of effective convnets [11] and corresponding AIP algorithms [36], some works [6, 10] have considered training convnets with AIPs, achieving robustness against AIPs to some extent. On the other hand, Graese *et al*. [7] argued that simple test time image processing, such as translation, Gaussian noise, blurring, and re-sizing, can equally neutralise the effect of AIPs, without having to retrain the convnet. In our case study, we include those image processing methods in the recogniser's strategy space.

Robust AIPs against classifiers. Sharif *et al*. [32] proposed a method for robustification by optimising an AIP against a set of images, rather than a single image. This approach was also suggested by Moosavi *et al*. [20] for generating *universal perturbations*. In our work, we consider optimising the AIP against a set of jittered versions of the target input. We will show empirically that this enables a targetted defense against image processing strategies.

AIP for identity obfuscation. This paper advocates the AIP as an effective and aesthetic means for disabling recognition. Previously Sharif *et al*. [32] also used adversarial optimisation to fool a person recogniser. Compared to their limited setup (fixed pose, fixed recognition strategy), our case study covers a large-scale social media setup with user-recogniser dynamics.

Person recognition task. Our case study considers the person recognition task in social media setup [5, 38, 27], as opposed to face recognition [9] (frontal faces, good lighting) or pedestrian re-identification [2, 1] (low resolution, fixed context). Social media photos capture subjects appearing in diverse range of viewpoints, poses, clothings, and events. Zhang *et al*. introduced PIPA [38], the first large-scale social media person recognition dataset and benchmark. Our empirical studies are built upon this dataset.

Person recognition models. Multiple researchers have proposed person recognition techniques in social media photos. Zhang *et al*. [38] proposed to combine cues from multiple body parts obtained by poselet detections. Oh *et al*. [27] greatly simplified [38] while achieving the state of the art performance. We build our recogniser model upon [27], possibly with more advanced network architectures. A concurrent work by Liu *et al*. [17] claims to have improved the method via metric learning objective. There exist other works [14, 28, 15], which exploit social media metadata.

3. User-Recogniser Game

This section provides a general framework for studying user-recogniser games. The framework provides a tool for systemising the path from a set of explicit assumptions on the players to game theoretical conclusions.

Our user-recogniser game framework is visualised in figure 2. The user U perturbs the original image x according to a strategy $i \in \Theta^u$, aiming to thwart recognition. The recogniser R processes the perturbed image $r_i(x)$ according to a strategy $j \in \Theta^r$, aiming to neutralise the effect of image perturbation. The resulting image $n_j(r_i(x))$ is passed to the model f to make a prediction. The game arises from the fact that each player does not know the opponent's strategy, although they do know each other's strategy space.

We introduce relevant game theoretical concepts and key theoretical results in §3.1 to help formalise the framework in §3.2. We discuss possible extensions in §3.3.

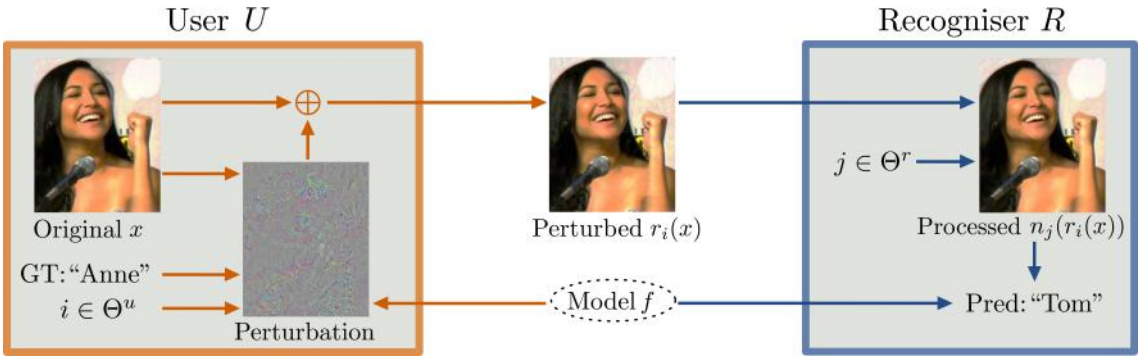

Figure 2: User-recogniser game on a single photo. Each player does not know the opponent's strategy. Orange (blue) arrows indicate actions taken by the user (recogniser). Information in the orange (blue) box is only available to the user (recogniser).

3.1. Two-Person Constant-Sum Games

We describe our system as a **two-person game** [26] consisting of two players, the user U and the recogniser R with designated **strategy spaces**, Θ^u and Θ^r.

As a result of each player committing to strategies $i \in \Theta^u$ and $j \in \Theta^r$ respectively, R receives a **payoff** of p_{ij}, the recognition rate; U then receives a payoff of $1 - p_{ij}$, the mis-recognition rate.

Game theory suggests that it is sometimes better to randomise the strategies. U can adopt a **mixed (random) strategy** $\theta^u = (\theta^u_i)_{i \in \Theta^u}$, defined as a distribution over the strategy space Θ^u, and similarly for R. With abuse of notation we write $p(\theta^u, \theta^r) := \sum_{i,j} \theta^u_i \theta^r_j p_{ij}$ for the expected payoff for R when the mixed strategies θ^u and θ^r are taken. The payoff for U is derived and defined as $\sum_{i,j} \theta^u_i \theta^r_j (1 - p_{ij}) = 1 - p(\theta^u, \theta^r) =: p'(\theta^u, \theta^r)$.

We say that a two-person game is a **constant-sum game** if the players' payoffs sum to a constant β independent of the strategies. In our case, the recognition and mis-recognition rates always sum to one ($\beta = 1$). A game is **finite** if the strategy spaces are finite. We have the following optimality theorem.

Theorem 1 (von Neumann [26], 1928)**.** *For a finite constant-sum game, there exist **optimal** or **minimax** mixed strategies $\theta^{u\star}$ and $\theta^{r\star}$ such that*

$$p(\theta^{u\star}, \theta^r) \leq p(\theta^{u\star}, \theta^{r\star}) \leq p(\theta^u, \theta^{r\star}) \quad \forall\, \theta^u, \theta^r \tag{1}$$

*where $v := p(\theta^{u\star}, \theta^{r\star})$ is the **value of the game**.*

Equation 1 implies that when R plays $\theta^{r\star}$, R is guaranteed to have a payoff of at least v, regardless of U's strategy; if U plays $\theta^{u\star}$, U is guaranteed to have a payoff of $1 - v$. In our scenario, this means that U's optimal strategy guarantees a certain mis-recognition rate, regardless of R's strategy.

U's optimal strategies can be obtained efficiently via linear programming that solves the following (R's optimal strategy can be found by swapping min and max):

$$\arg\min_{\theta^u} \max_{\theta^r} \sum_{i,j} \theta^u_i \theta^r_j p_{ij} \quad \text{s.t. } \theta^u,\ \theta^r \text{ are distributions.} \tag{2}$$

If U has knowledge on R's strategy $\bar{\theta}^r$, then U can take advantage of this knowledge. U can optimise her strategy given $\bar{\theta}^r$ to attain a payoff of $\max_{\theta^u} p'(\theta^u, \bar{\theta}^r) \geq p'(\theta^{u\star}, \bar{\theta}^r) \geq p'(\theta^{u\star}, \theta^{r\star}) = 1 - v$, a potentially better payoff than the no-knowledge scenario $1 - v$. However, if R's strategy is optimal $\bar{\theta}^r = \theta^{r\star}$, then the knowledge does not bring improvement for U: $\max_{\theta^u} p'(\theta^u, \theta^{r\star}) = 1 - v$.

In reality, not all players play optimally either due to the lack of knowledge (*e.g.* on the opponent's strategy space), or due to pure irrationality. We refer to such a player as an **irrational player**. Our discussion above implies:

Corollary 1. *If U knows R's strategy $\bar{\theta}^r$, and if it is suboptimal, then U can enjoy a better payoff than $1 - v$.*

3.2. Components of the User-Recogniser Game

We specify the payoffs, strategy spaces, and information allowed for the user U and the recogniser R.

Test data. We assume that the test data are distributed according to $(\hat{x}, \hat{y}) \sim D$. This dataset is the source of private information that the two players compete for.

Fixed model. We assume that U and R use a fixed model f (*e.g.* a publicly available model). This is a reasonable assumption, as U and R often would not have resources to train modern convnets.

Known model. Each player is aware that the opponent uses f. This may be unrealistic, but provides a good starting point. Relaxation of this assumption is discussed in §3.3.

Payoff. When the players commit to strategies $i \in \Theta^u$ and $j \in \Theta^r$, R's payoff is the recognition rate on the test set:

$$p_{ij} = \mathop{\mathbb{P}}_{(\hat{x},\hat{y}) \sim D} \left[\arg\max_y f^y \left(n_j \left(r_i \left(\hat{x} \right) \right) \right) = \hat{y} \right] \tag{3}$$

where f^y denotes the model prediction score for class y. U receives the payoff $1 - p_{ij}$, the mis-recognition rate.

User's strategy space Θ^u. We consider additive perturbations such that for an input x,

$$r_i(x) = x + t(x), \qquad ||t(x)||_2 \leq \epsilon \tag{4}$$

for some constant $\epsilon > 0$. When ϵ is small enough, the perturbation is nearly invisible to human eyes (see figure 3). These perturbations are frequently referred to as *adversarial image perturbations* (AIPs). We discuss existing AIPs and our novel variants in §4.

Recogniser's strategy space Θ^r. R aims to neutralise the adversarial effect of AIPs. Although some works have suggested re-training the model with AIPs, demonstrating certain degree of robustification [6, 10], Graese *et al.* [7] has argued that simple image processing can already neutralise the AIP effects cheaply and effectively. They have demonstrated that on MNIST, translation (T), Gaussian additive noise (N), blurring (B), and cropping & re-sizing (C) have improved the recognition rate from 0% (post-AIP) to 68%, 58%, 65%, and 76%, respectively. In our case study, we will include these transformations in Θ^r. In §3.3, we will discuss about expanding strategy spaces.

Known strategy spaces. The strategy spaces for each player (Θ^u and Θ^r) are known to each other, while the chosen strategies are not known.

Multiple recognisers. U may encounter a set of recognisers not all of which are malicious. For example, U uploads her personal photos to a cloud service with a recognition system R_1; she wants an AIP that enables a successful recognition by R_1 but disables recognition by a malicious system R_2. We propose an approach for generating *selective* AIPs in §4.2 and confirm their existence in §5.5. From a theoretical standpoint, the existence of selective AIPs attest to the diversity of possible AIP patterns, in line with the existence of *universal perturbations* [20].

3.3. Extensions

In the previous section, we have introduced the user-recogniser game framework with particular assumptions explored in this paper. In this section, we show that the framework can be extended beyond this setup.

Unknown models. Many AIP techniques assume a full knowledge on the model f, but the computation of *black-box* AIPs is another active research field [29, 30, 24, 16]; U can potentially adopt these methods.

Non-constant sum. If U and R assign different weights to different test samples, then the payoffs may not sum to 1. For such non-constant sum games, there exist *Nash equilibrium* strategies for each player [25]. The optimal strategy and payoff analyses are still possible.

Non-additive AIPs. The framework allows r_i to be any function that induces invisible changes on the image. Current restriction to equation 4 rules out *e.g.* one-pixel translation of the whole image. Most, if not all, prior work on AIP is done in the additive setup. Crafting non-additive AIP would be interesting future work.

Non-fixed models. R with enough computational resources may re-train the model f with AIPs. One option to expand our framework to such a setup would be to incorporate the model parameters in Θ^r. Brückner *et al.* [3] have studied this setup, but have assumed convex loss functions. Understanding games with continuous strategy spaces and non-convex payoffs (*e.g.* convnet losses) is an open question both for computer vision and game theory research.

Unknown strategy spaces. The exact possible set of strategies may not be known to the opponent. With improving technologies, the respective strategy spaces may even grow over time. The framework cannot do much about the unknown strategies, but can adaptively expand the strategy spaces according to technological developments.

4. Adversarial Image Perturbation Strategies

This section reviews existing adversarial image perturbation (AIP) algorithms that use first-order optimisation schemes, and proposes our novel variants.

We compute AIPs as additive transformations with L_2 norm constraints (equation 4). Computation of AIP can be formulated as a loss *maximisation* problem

$$\max_t \mathcal{L}(f(x+t), y) \qquad \text{s.t. } ||t||_2 \leq \epsilon \tag{5}$$

where x is the input image and y is the ground truth label; the loss function $\mathcal{L}$ is to be specified.

4.1. Existing AIP methods

Depending on the loss function $\mathcal{L}$ and the optimisation algorithm, we recover most of the existing AIP methods such as Fast Gradient Vector [31], Fast Gradient Sign [6], Basic Iterative [12], and DeepFool [21]. The *universal perturbations* introduced by Moosavi *et al.* [20] can also be seen as a special case of equation 5 where the loss is computed over the entire test set and the perturbation t is shared across images. See table 1 for the summary.

Fast Gradient Vector (`FGV`) [31]. `FGV` adopts the softmax-log loss $\mathcal{L} = -\log \hat{f}^y$ in equation 5, solving it via one-step gradient ascent: $t^\star = -\gamma \nabla \mathcal{L}(x)$ for some constant $\gamma > 0$.

Fast Gradient Sign (`FGS`) [6]. `FGS` is identical to `FGV`, except that $\nabla \mathcal{L}(x)$ is replaced with $\text{sign}(\nabla \mathcal{L}(x))$.

Gradient Ascent (`GA`). This is a multi-step variant of `FGV`. Perturbation is initialised at $t^{(0)} = 0$. Gradient ascent is performed on the loss function iteratively: $t^{(m+1)} = t^{(m)} - \gamma \nabla \mathcal{L}(x + t^{(m)})$ for $m = 0, \cdots, K$ for some fixed step size $\gamma > 0$ and maximal number of iterations $K \geq 1$.

Variants	Loss $\mathcal{L}$	Stopping condition	Step size
FGS[6]	$-\log \hat{f}^y$	1 iteration	Fixed
FGV[31]	$-\log \hat{f}^y$	1 iteration	Fixed
BI[12]	$-\log \hat{f}^y$	K iterations	Fixed
GA	$-\log \hat{f}^y$	K iterations	Fixed
DF[21]	$f^{y^c} - f^y$	K it.$\vee$ fooled	Adaptive
GAMAN	$f^{y^\star} - f^y$	K iterations	Fixed

Table 1: Conceptual differences among AIP methods. $f^{y'}$ is the model score for class y', and $\hat{f}$ denotes the softmax output of f. y is the ground truth label, and $y^\star$ is the most likely label among wrong ones. y^c is the label with the closest linearised decision boundary.

Basic Iterative (BI) [12]. BI is identical to GA, except that $\nabla\mathcal{L}(x)$ is replaced with sign $(\nabla\mathcal{L}(x))$.

DeepFool (DF) [21]. DF algorithm solves the objective:

$$\min_t ||t||_2 \quad \text{s.t.} \quad \arg\max_y f^y(x+t) \neq y \tag{6}$$

which finds the minimal perturbation such that the prediction is wrong. Although the objective is different, we show that the DF algorithm can also be seen as a first-order method solving equation 5 for some loss function.

DF first finds the class with the nearest decision hyperplane, denoted by c. To simplify the search, c is found on the linear approximation of f around x (tangent function). The normal vector to the decision hyperplane is given by $\nabla f^c - \nabla f^y$. At each iteration, the algorithm computes the minimal step size along this direction to reach the decision hyperplane. Since f is not linear, the algorithm may need more than one iterations to cross the decision hyperplane.

We observe that if we set the loss function as $\mathcal{L} = f^c - f^y$ the gradient ascent direction matches the DF step directions $\nabla f^c - \nabla f^y$. We thus regard DF as a gradient ascent algorithm with each step size minimised to just induce a wrong prediction.

Projection and clipping. The norm constraint $||\cdot||_2 \leq \epsilon$ as well as RGB value constraint to $[0, 255]$ must be enforced on the solution. [16, 12] suggest applying projections after each iteration. We follow this practice. For BW images, we average the gradients for each RGB channel.

4.2. Our AIP methods

As we will demonstrate in §5.2, the above approaches are fragile to simple image processing techniques. We propose novel AIP approaches here, focusing on robustness.

[1] *Gaman* is a Zen Buddhist term for *endurance*.

Gradient Ascent – Maximal Among Non-GT (GAMAN[1]). Even if the prediction label is changed by the AIP, this would not be robust if the perturbed input is still close to the decision boundary. DeepFool (DF) is not expected to be robust, as it stops iterations as soon as the decision boundary is reached. On the other hand, DF guides the solution to the closest decision boundary; if we let DF iterate beyond the decision boundary with a fixed step size with fixed number of iterations, the solution is likely to proceed more deeply into the territory of the wrong label, improving robustness.

This motivates our GAMAN variant. Instead of the costly computation of c at each iteration, we approximate $c \approx y^\star := \arg\min_{y' \neq y} f^{y'}$, the most likely prediction among wrong labels. We set the loss function as $\mathcal{L} = f^{y^\star} - f^y$, and perform gradient ascent with a fixed step size γ for K iterations. This approach is similar but different from the impersonation AIPs previously considered [32, 16], which drive the solution to a fixed impersonation target $\bar{y}$. In contrast, $y^\star$ may change during the iterations.

Vaccination against image processing. The above methods maximise classification loss functions with respect to a fixed recogniser. For countering an AIP-neutralising image processing technique n_j, we consider including the image processing step in the loss function: $\mathcal{L}(n_j(x+t))$. Any first-order method considered above can be used, as long as n_j is differentiable. If the processing function is random, we average the gradients from multiple samples. We refer to this technique as *vaccination*. Note that this technique is complimentary to the above mentioned methods.

Selective AIPs. We present another complimentary technique for generating AIPs targetted to a selected subset of recognisers. To avoid recognition from $\mathcal{M}$ while authorising $\mathcal{B}$ to recognise, we propose to maximise a mixed loss

$$\sum_{k\in\mathcal{M}} \lambda_k \mathcal{L}_k - \sum_{k'\in\mathcal{B}} \lambda_{k'} \mathcal{L}_{k'} \tag{7}$$

with $\lambda_k, \lambda_{k'} > 0$.

5. Empirical Studies

We have set up a game theoretical framework to study the dynamics between the user U and the recogniser R. In particular, previous adversarial image perturbation (AIP) techniques are studied, and new variants are proposed.

In this section, we present a case study of the framework on *person recognition*. Before presenting the game theoretical analysis, we evaluate the performance of existing and newly proposed AIP techniques (§5.2), and the effectiveness of R's image processing strategies Θ^r (§5.3). The full game is introduced (§5.4) after specifying U's strategy space; we study this system in depth. Finally, we show results on the recogniser-selective AIPs (§5.5).

	Perturbation	AlexNet	VGG	Google	ResNet
	None	83.8	86.1	87.8	91.1
Image Proc.	Noise	≥83	≥85	≥87	≥90
	Blur	≥82	≥85	≥86	≥90
	Eye Bar	≥81	≥84	≥84	≥87
1-Iter. AIP	FGS[6]	23.6	16.0	5.9	20.2
	FGV[31]	13.3	11.5	4.6	20.0
K-Iter. AIP	BI[12]	1.2	0.5	0.0	0.0
	GA	0.2	0.0	0.0	0.0
	DF[21]	0.0	0.0	0.0	0.0
	GAMAN	0.0	0.0	0.0	0.0

Table 2: Recognition rates after image perturbation. In all methods, the perturbation is restricted to $||\cdot||_2 \leq 1000$. For the baseline image processing perturbations, we only report lower bounds (denoted $\geq \cdot$).

5.1. Dataset and Experimental Setup

Dataset. We build our analysis upon the PIPA (People In Photo Albums) [38], a large-scale dataset of social media photos crawled from Flickr. We use the $\texttt{val}_1$ subset of PIPA, consisting of 4820 instances of 366 identities ($\texttt{val-original split}_1$ in [27] terminology) as the test set. We assume that the user uploads cropped head images to social media; PIPA provides the GT head boxes.

Person recogniser. The person recognition model f is built on a state of the art framework [27]. It first trains a convnet for the person recognition task on a large database of random identities; it then tunes the final classification layer to the test identities using about ten examples per identity. In our case, we have used the $\texttt{val}_0$, which is of the same size and set of identities as as $\texttt{val}_1$. While [27] only considered AlexNet [11], we also consider VGG [33], GoogleNet [35], and ResNet152 [8]. They show better recognition rates (table 2).

Evaluation. We evaluate payoffs for R in terms of the ratio of correctly identified instances in the test set. The payoff for U is 1 minus R's payoff. In all the tables, R is the column player and U is the row player. For each column (row), U's (R's) optimal strategy is marked orange (blue).

5.2. Comparison of Perturbation Methods

AIP parameters. We set $\epsilon = 1000$ in all our experiments, unless stated otherwise. For GoogleNet input 224×224, this corresponds to 2% of pixels perturbed by 1/256. For Gradient Ascent (GA) and Basic Iterative (BI) the step size γ is set to 10^4; for GAMAN, 5×10^3. We set the maximal number of iterations $K = 100$, determined such that the norm reaches $\epsilon = 1000$ in K iterations for most test samples.

Baseline perturbation methods. We consider three commonly used obfuscation types: noise, blur, and eye bar. Noise adds iid Gaussian noise of variance σ^n; blur performs convolution with a Gaussian kernel of size σ^b; eye bar puts a gray horizontal bar of thickness σ^e on the upper $\frac{1}{3}$ location. They incur large L_2 distances (> 1000) from the original image even with small σ^n, σ^b, and σ^e. In table 2, we report the *lower bounds* on the recognition rates at $||\cdot||_2 = 1000$ by computing the rates at some $||\cdot||_2 > 1000$.

AIP performance. We first evaluate all the considered AIP methods against all network variants. Table 2 shows the results. We observe that noise, blur, and eye bar have nearly no impact on the recognition performance for small L_2 perturbations. AIP variants show better obfuscation performances. Vanilla gradient overall gives better obfuscation than signed versions; on AlexNet Fast Gradient Vector (FGV) reduces the recognition rate to 13.3, compared to 23.6 for Fast Gradient Sign (FGS); the multi-iteration analogues show similar behaviours with Gradient Ascent (GA) achieving 0.2 compared to 1.2 by Basic Iterative (BI). Finally, we observe that the DeepFool (DF) and GAMAN (§4.2) are very effective, pushing the recognition rates down to zero.

Network performance. Comparing architectures, we observe that AlexNet is surprisingly robust to AIPs compared to more recent architectures. GoogleNet, for example, performs better than Alexnet without AIPs (83.8 vs 87.8); when FGS is used, AlexNet performs 23.6 while GoogleNet performs 5.9. When multi-iteration AIPs are used, the architectural choice does not have a significant impact. We opt for GoogleNet in the next experiments; it is reasonably performant, while being much faster than ResNet.

5.3. Robustness of AIPs

Basic processing Proc. Even before R's image processing strategies take place, the perturbed image needs to be (1) re-sized to the original image (from the network input sizes) and (2) quantised to integer values (*e.g.* 24-bit true color). We denote the above two basic processing steps as Proc.

Image processing strategies Θ^r. We fully specify R's strategy space for our case study. Following Graese *et al.* [7], we consider $\Theta^r = \{\text{Proc}, \text{T}, \text{N}, \text{B}, \text{C}, \text{TNBC}\}$. Proc is the basic processing described above, and all the other strategies are applied over Proc. T is translation by a random offset within 10% of the image side lengths. N adds iid Gaussian noise with variance $\sigma^2 = 10^2$. B blurs with Gaussian kernel of width chosen from $\{1, 3, 5, 7, 9\}$ uniformly at random. C crops with a random offset within 10% of the image side lengths and re-sizes back to the original. For each strategy, the recogniser ensembles the scores from five random samples. We also consider the combination of all four (TNBC). It runs the model four times on each processed image and once on the original; the scores are then averaged.

Perturb	∅	Proc	T	N	B	C	TNBC
None	87.8	87.8	87.6	64.0	81.2	85.4	87.3
BI[12]	0.0	8.3	15.8	16.8	28.6	27.4	17.6
GA	0.0	8.6	13.2	14.1	28.4	23.7	16.4
DF[21]	0.0	51.8	75.6	56.5	72.5	76.9	75.5
GAMAN	0.0	4.0	6.6	15.0	22.2	16.7	9.9

Table 3: Robustness analysis of AIPs on GoogleNet. AIPs are restricted to to $||\cdot||_2 \leq 1000$. Proc indicates the resizing and quantisation needed to convert AIP outputs to image files. (T, N, B, C) = (Translate, Noise, Blur, Crop).

	Recogniser Θ^r					
User Θ^u	Proc	T	N	B	C	TNBC
GAMAN	4.0	6.6	15.0	22.2	16.7	9.9
/T	2.5	2.3	11.6	18.5	7.2	4.9
/N	5.8	7.6	4.6	23.6	16.6	9.1
/B	0.4	0.8	8.6	5.8	3.1	1.4
/C	2.6	2.2	11.8	18.1	3.4	4.3
/TNBC	0.7	0.9	5.2	9.5	3.2	2.0

Table 4: Recogniser's payoff table p_{ij}, $i \in \Theta^u$ and $j \in \Theta^r$. The user's payoff is given by $100 - p_{ij}$.

Robustness of AIPs. Table 3 shows the recognition rates for the GoogleNet when R's processing strategies are present. While the multi-iteration AIPs induce zero recognition rates without any processing, Proc already exhibits powerful neutralisation effects: recognition rates for Gradient Ascent (GA) and DeepFool (DF) jump from zero to 8.6 and 51.8, respectively. The instability of DF is due to early stopping (§4.1). The processing strategies by R further increase recognition rates. Blurring B and cropping C strategies prove to be more harmful to AIPs than translation T and noise N in general. Comparing AIP-wise, we show that our novel variant GAMAN (§4.2) dominates other methods against all processing strategies but N; GA performs better in that case, but only by a small amount (14.1 versus 15.0). Subsequent analyses are built on GAMAN.

Qualitative. Qualitative examples of the methods are shown in figure 3. The images and the prediction results are after Proc. GA and GAMAN reliably induces misidentification without sacrificing aesthetics compared to blurring.

5.4. User-Recogniser Games

Vaccination strategies Θ^u. In response to the processing strategies by the recogniser R, the user U may vaccinate the AIP against expected processing types (§4.2). We consider six variants $\Theta^u = \{$GAMAN, /T, /N, /B, /C, /TNBC$\}$. We use slash / to indicate vaccination on GAMAN. For /T, /N, /B, /C, gradients from 5 random function samples are averaged at each iteration. The combination strategy /TNBC averages 4 gradients from individual methods and 1 original gradient, resulting in the same number of gradient computations for all vaccination variants.

Is vaccination helpful? Table 4 shows the recognition rates of GoogleNet for combinations of discussed processing and vaccination strategies. We observe indeed that each vaccination type makes the vanilla AIP GAMAN more robust against the respective processing type: for B the rate drops from 22.2 to 5.8. /B is the most effective strategy for U against all processing strategies except for N. For N, the corresponding vaccination /N yields the best payoff for U. We conjecture this is because the noise N results in high frequency patterns while the others smooth the output. We observe, finally, that the combined vaccination /TNBC cannot prepare AIP against all processing types most effectively; given a budget on the number of gradient computations, it is hard to be good at everything.

Optimal deterministic strategy. We can regard table 4 as the payoff table p_{ij} for R for strategies $i \in \Theta^u$ and $j \in \Theta^r$. Let's first assume that the players only choose fixed strategies. Then, solving equation 2 with determinism constraints $\theta^u_i, \theta^r_j \in \{0, 1\}$ yields U's optimal strategy as /B with a privacy guarantee of at most 8.6 recognition rate.

Optimal random strategy. Game theory suggests that it is sometimes better to randomise strategies. Solving equation 2 without the integral constraints yield the optimal solutions for U and R as $\theta^{u\star} = (/\text{B} : 61\%, /\text{TNBC} : 39\%)$ and $\theta^{r\star} = (\text{N} : 52\%, \text{B} : 48\%)$, respectively. Playing $\theta^{u\star}$ guarantees U to allow at most 7.3 recognition rate, an improved privacy guarantee than the deterministic case, 8.6.

Knowledge on R's strategy. As discussed in §3.1, having knowledge on R's strategy can improve the payoff bound for U, if R does not play the optimal strategy. Let us consider two possible non-optimal strategies played by R. (1) If R commits to B, U's optimal strategy is the minimal row in the column B: /B, with recognition rate 5.8. (2) If R randomises uniformly over Θ^r, U's optimal strategy is the minimal row over the column average: /B with recognition rate 3.4. In both cases, U enjoys lower recognition rates.

Limited knowledge on Θ^r. Assume that U is not aware of all possible technologies that R has at hand. For example, the strategy N is not known to U. Then, U's apparent optimal solution is $(/\text{B} : 100\%)$, which she thinks will guarantee her at most 5.8 recognition rate. R can then attack U with N, incurring 8.6 recognition rate. Limited knowledge on the opponent's strategy space does hurt.

5.5. Selective AIPs

We assume that U wants to avoid identification by a set of malicious recognisers $\mathcal{M}$, while authorising identification

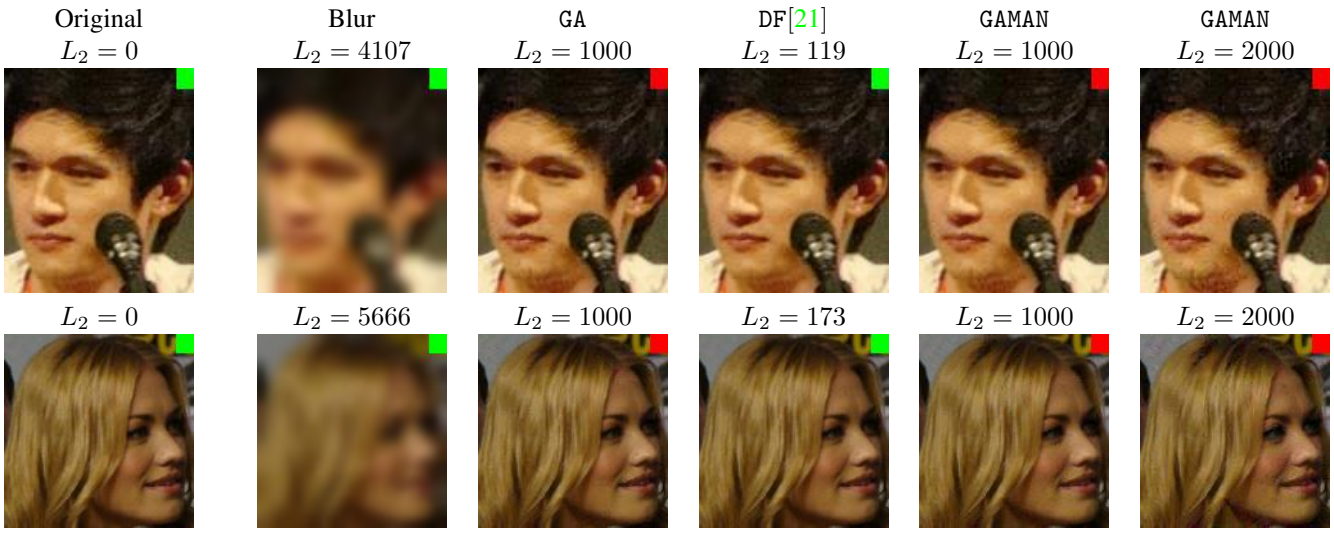

Figure 3: Perturbed images after Proc and the corresponding predictions (green for correct, red for wrong). GA and GAMAN reliably confuse the classifier at almost no cost on the aesthetics. At $L_2 = 2000$, GAMAN does show small artifacts.

Setup			$\mathcal{M}$ averaged		$\mathcal{B}$ averaged	
$\mathcal{M}$	$\mathcal{B}$	L_2	w/o AIP	w/ AIP	w/o AIP	w/ AIP
{G}	$\emptyset$	1000	87.8	4.0	-	-
{G}	{A}	1000	87.8	8.7	83.8	97.9
{A,R}	{V,G}	1000	87.4	17.7	87.0	97.7
{A,R}	{V,G}	2000	87.4	3.8	87.0	97.8

Table 5: Selective AIPs. AIPs are crafted to confuse $\mathcal{M}$ leaving $\mathcal{B}$ intact. [A,V,G,R] = [AlexNet, VGG, GoogleNet, ResNet152]. GAMAN has been used in all experiments. Reported performances are after Proc.

by benign ones $\mathcal{B}$. We set up the experiments in table 5. We include the GAMAN performance on GoogleNet as a baseline (first row). We solve equation 7 with $\lambda_k = 1$ for all $k \in \mathcal{M} \cup \mathcal{B}$ to generate selective AIPs.

When $\mathcal{M} = \{\text{GoogleNet}\}$ and $\mathcal{B} = \{\text{AlexNet}\}$, the generated AIP incurs mere 8.7 identification for $\mathcal{M}$ (after Proc), while allowing $\mathcal{B}$ to identify 97.9 percent. We thus confirm the selectivity. However, this comes at the cost of increased recognition rate for $\mathcal{M}$ (8.7), compared to when AIP only had to confuse $\mathcal{M}$ (4.0).

We also consider the multi-$\mathcal{M}$, multi-$\mathcal{B}$ case given by $\mathcal{M} = \{\text{AlexNet, ResNet}\}$ and $\mathcal{B} = \{\text{VGG, GoogleNet}\}$. The average performance is 17.7 for $\mathcal{M}$, and 97.7 for $\mathcal{B}$, post Proc. Selectivity thus works for multiple models, but again the recognition rates for $\mathcal{M}$ are quite high (17.7). We remark that by increasing the budget on perturbation size from 1000 to 2000, we can still attain a lower rate: 3.8.

The existence of selective AIPs is not only of practical but also of theoretical interest. They show that the space of AIPs is diverse enough to accommodate patterns that simultaneously hamper and assist recognition.

6. Discussion & Conclusion

Game theoretical approach. Game theory is a tool for wading through uncertainties in players' choices, providing payoff guarantees independent of the opponent's strategies. Game theory also suggests that if there is no single technology which best copes with all possible adversarial technologies, it is better to randomise existing techniques.

As discussed in §3.3, the game theoretical framework introduced in this paper can be extended to other setups, where less resource constraints are placed on each player. This paper serves as a first step towards the promising research direction of analysing the user-recogniser dynamics.

Conclusion. In this work, we have constructed a game theoretical framework to study a system with two players, user U and recogniser R, with antagonistic goals (dis-/enable recognition). We have examined existing and new adversarial image perturbation (AIP) techniques for U. As a case study of the framework, we have presented a game theoretical analysis of the privacy guarantees for a social media user, assuming strategy spaces that include the state of the art AIPs and person recognition techniques.

Acknowledgement. This research was supported by the German Research Foundation (DFG CRC 1223). We thank Dr Yun Kuen Cheung for exciting discussions on the Game Theory and comments on the manuscript. We also thank Dr Mykhaylo Andriluka and Tribhuvanesh Orekondy for helpful comments on the paper.

References

[1] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. In *IVC*, 2014.

[2] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *BMVC*, 2009.

[3] M. Brückner, C. Kanzow, and T. Scheffer. Static prediction games for adversarial learning problems. In *JMLR*, 2012.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[5] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[7] A. Graese, A. Rozsa, and T. E. Boult. Assessing threat of adversarial examples on deep neural networks. In *ICMLA*, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, UMass, 2007.

[10] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári. Learning with a strong adversary. *CoRR*, abs/1511.03034, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[12] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *CoRR,abs/1607.02533*, 2016.

[13] G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *JLMR*, 2003.

[14] H. Li, J. Brandt, Z. Lin, X. Shen, and G. Hua. A multi-level contextual model for person recognition in photo albums. In *CVPR*, 2016.

[15] Y. Li, G. Lin, B. Zhuang, L. Liu, C. Shen, and A. van den Hengel. Sequential person recognition in photo albums with a recurrent network. In *CVPR*, 2017.

[16] Y. Liu, X. Chen, C. Liu, and D. X. Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

[17] Y. Liu, H. Li, and X. Wang. Learning deep features via congenerous cosine loss for person recognition. *CoRR,abs/1702.06890*, 2017.

[18] R. McPherson, R. Shokri, and V. Shmatikov. Defeating image obfuscation with deep learning. *CoRR,abs/1609.00408*, 2016.

[19] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *WSDM*, 2010.

[20] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

[21] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[22] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *SP*, 2009.

[23] A. Narayanan and V. Shmatikov. Myths and fallacies of personally identifiable information. *CACM*, 2010.

[24] N. Narodytska and S. P. Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *CoRR,abs/1612.06299*, 2016.

[25] J. F. Nash. Equilibrium points in n-person games. *PNAS*, 1950.

[26] J. v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 1928.

[27] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *ICCV*, 2015.

[28] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition; privacy implications in social media. In *ECCV*, 2016.

[29] N. Papernot, P. McDaniel, and I. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR,abs/1605.07277*, 2016.

[30] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples. In *ASIACCS*, 2017.

[31] A. Rozsa, E. M. Rudd, and T. E. Boult. Adversarial diversity and hard positive generation. In *CVPRW*, 2016.

[32] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *SIGSAC*, 2016.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[34] Q. Sun, M. Fritz, and B. Schiele. A domain based approach to social relation recognition. In *CVPR*, 2017.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[37] M. J. Wilber, V. Shmatikov, and S. Belongie. Can we still avoid automatic face detection? In *WACV*, 2016.

[38] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR*, 2015.

[39] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *WWW*, 2009.

Supplementary Materials

A. Contents

The supplementary materials contain auxiliary experiments for the empirical analyses in the main paper. In particular, we include:

- Score loss for adversarial image perturbation (AIP).
- AIP performance at different L_2 norms.
- Experiments for the non-GoogleNet architectures.
- More qualitative results.

As in the main paper, we mark the optimal entry in each column (row) for the user (recogniser) with orange (blue).

B. Score Loss for AIPs

In the main paper, we have reviewed variants of AIPs according to the loss functions and the optimisation algorithms. Algorithms `FGV`, `FGS`, `BI`, and `GA` use the softmax-log loss $-\log \hat{f}^y$. The DeepFool (`DF`) and our `GAMAN` variants use the difference of two scores (*e.g.* $f^{y^\star} - f^y$). This section includes an auxiliary analysis for the effect of the loss type: softmax-log loss $-\log \hat{f}^y$ versus score loss $-f^y$. We denote the score loss analogues with the suffix `-S` (*e.g.* `FGS-S`). We also include `FGMAN` (Fast Gradient – Maximal Among Non-GT), the single iteration analogue of `GAMAN`, for completeness. See table 6 for a summary.

The corresponding empirical performances are shown in table 7 and 9. Since single-iteration AIPs are significantly outperformed by the multi-iteration AIPs, we have focused on the latter in the main paper, and so do we here. In table 7, we observe that the choice of the loss function does not make much difference. Table 9 further supports this view against image processing techniques, although the softmax-log loss does perform marginally better.

C. AIP Performance at Different L_2 Norms

In the main paper, we have used the L_2 norm constraint $\epsilon = 1000$ as the default choice. In this section, we examine the behaviour of AIP performance at varying ϵ values.

See figure 4 for the plot. The performances are post-Proc (§5.3). We fix the step size to $\gamma = 10^4$ (5×10^3 for `GAMAN`), and the maximal number of iterations to $K = 100$; we choose the norm constraint ϵ from $\{100, 200, 500, 1000, 2000\}$. The norm of the resulting AIP is upper bounded by ϵ, but may not necessarily be exactly ϵ. The average norm across the test set is plotted.

We observe that the AIP variants are much more effective than Noise, Blur, or Eye Bar, achieving the same degree of obfuscation at $1 \sim 2$ orders of magnitude smaller perturbations. At the same norm level, the multi-iteration variants (`BI`,`GA`) are more effective than the single-iteration analogues (`FGS`,`FGV`). Taking gradient signs decreases the

Variants	Loss $\mathcal{L}$	Stopping condition	Step size
`FGS`[6]	$-\log \hat{f}^y$	1 iteration	Fixed
`FGV`[31]	$-\log \hat{f}^y$	1 iteration	Fixed
`FGS-S`	$-f^y$	1 iteration	Fixed
`FGV-S`	$-f^y$	1 iteration	Fixed
`FGMAN`	$f^{y^\star} - f^y$	1 iteration	Fixed
`BI`[12]	$-\log \hat{f}^y$	K iterations	Fixed
`GA`	$-\log \hat{f}^y$	K iterations	Fixed
`BI-S`	$-f^y$	K iterations	Fixed
`GA-S`	$-f^y$	K iterations	Fixed
`DF`[21]	$f^{y^c} - f^y$	K it.$\vee$ fooled	Adaptive
`GAMAN`	$f^{y^\star} - f^y$	K iterations	Fixed

Table 6: Extended version of table 1 in the main paper; additional methods are denoted as gray cells. $f^{y'}$ is the model score for class y', and $\hat{f}$ denotes the softmax output of f. y is the ground truth label, and $y^\star$ is the most likely label among wrong ones. y^c is the label with the closest linearised decision boundary. $\tilde{y}$ is the least likely label.

	Perturbation	AlexNet	VGG	Google	ResNet
	None	83.8	86.1	87.8	91.1
Image Proc.	Noise	$\geq$83	$\geq$85	$\geq$87	$\geq$90
	Blur	$\geq$82	$\geq$85	$\geq$86	$\geq$90
	Eye Bar	$\geq$81	$\geq$84	$\geq$84	$\geq$87
1-Iter. AIP	`FGS`[6]	23.6	16.0	5.9	20.2
	`FGV`[31]	13.3	11.5	4.6	20.0
	`FGS-S`	27.8	6.2	1.0	4.3
	`FGV-S`	21.0	5.5	3.5	8.0
	`FGMAN`	4.4	3.9	2.8	11.5
K-Iter. AIP	`BI`[12]	1.2	0.5	0.0	0.0
	`GA`	0.2	0.0	0.0	0.0
	`BI-S`	1.2	0.3	0.0	0.0
	`GA-S`	0.2	0.0	0.0	0.0
	`DF`[21]	0.0	0.0	0.0	0.0
	`GAMAN`	0.0	0.0	0.0	0.0

Table 7: Extended version of table 2 in the main paper; new entries are denoted as gray cells. Recognition rates after image perturbation. In all methods, the perturbation is restricted to $||\cdot||_2 \leq 1000$. For the baseline image processing perturbations, we only report lower bounds (denoted $\geq \cdot$).

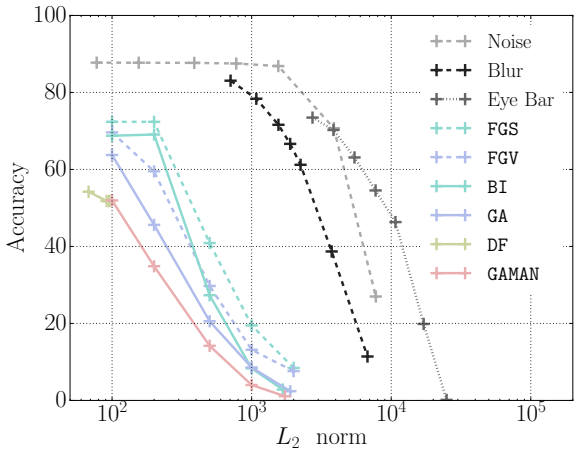

Figure 4: GoogleNet accuracy after various perturbations methods at different L_2 norms. All results are after Proc.

obfuscation performance at small L_2 norms ($\leq$ 1000), but they converge to a similar performance at $\epsilon = 2000$. DeepFool (DF) outputs have small norms $\leq$ 100 due to early stopping. Our variant GAMAN performs best across all norm levels, achieving nearly zero recognition at $\epsilon = 2000$.

D. Non-GoogleNet Experiments

In the main paper, we have focused on the GoogleNet results for the AIP robustness analysis and the game theoretic studies (table 3 and 4). We extend the experiments to AlexNet, VGG, and ResNet152.

D.1. Robustness Analysis

See table 9 for the robustness analyses for all four networks. We confirm here again that GAMAN shows overall best robustness, across image processing techniques (Proc, T, N, B, C, and TNBC), across architectures. For AlexNet and ResNet, cropping (C) is the most powerful neutralisation, while for VGG and GoogleNet blurring (B) is. We observe that the effects are particularly strong for ResNet; C boosts the performance from 0.0 to 31.8 against GAMAN.

D.2. Game Analysis for Various Networks

See table 10 for the payoff tables for all four networks. We summarise the optimal user strategy $\theta^{u\star}$ and the corresponding guarantee on the recognition rate in table 8. Note that against all but AlexNet architecture, the optimal strategy $\theta^{u\star}$ is given as a mixture of /B and /TNBC.

E. Additional Qualitative Results

We include more qualitative results (equivalent to figure 3 in the main paper). See figures 5, 6, 7, 8.

Network	Optimal Strategy $\theta^{u\star}$	Bound on Rec. Rate
AlexNet	(/B : 100%)	$\leq$6.4
VGG	(/B : 86%, /TNBC : 14%)	$\leq$4.9
GoogleNet	(/B : 61%, /TNBC : 39%)	$\leq$7.3
ResNet	(/B : 31%, /TNBC : 69%)	$\leq$8.5

Table 8: Optimal strategies and the corresponding guaranteed upper bounds on the recognition rate for different networks. We write $\leq \cdot$ to denote the upper bound.

AlexNet

Perturb	∅	Proc	T	N	B	C	TNBC
None	83.8	83.8	83.7	77.8	78.7	80.1	83.9
BI[12]	1.2	10.0	29.7	20.8	26.6	34.3	23.3
GA	0.2	4.8	13.6	11.6	17.7	17.8	12.2
BI-S	1.2	10.1	31.2	21.0	27.2	35.7	23.3
GA-S	0.2	5.0	15.4	12.6	19.0	19.3	12.8
DF[21]	0.0	62.1	76.5	68.5	69.4	75.0	74.7
GAMAN	0.0	1.4	6.4	9.2	13.5	12.3	5.6

VGG

Perturb	∅	Proc	T	N	B	C	TNBC
None	86.1	86.1	84.8	77.2	81.5	84.1	85.8
BI[12]	0.5	6.8	11.1	18.1	23.2	16.8	14.4
GA	0.0	4.2	5.5	11.2	17.2	10.2	8.2
BI-S	0.3	7.1	11.2	19.2	23.8	17.3	14.3
GA-S	0.0	4.8	5.9	11.9	18.6	11.3	8.8
DF[21]	0.0	53.3	66.3	65.9	69.4	69.2	71.4
GAMAN	0.0	1.6	2.1	8.5	11.8	5.6	3.5

GoogleNet

Perturb	∅	Proc	T	N	B	C	TNBC
None	87.8	87.8	87.6	64.0	81.2	85.4	87.3
BI[12]	0.0	8.3	15.8	16.8	28.6	27.4	17.6
GA	0.0	8.6	13.2	14.1	28.4	23.7	16.4
BI-S	0.0	8.8	17.2	17.7	29.3	28.8	18.8
GA-S	0.0	9.1	14.9	15.2	29.3	25.5	18.0
DF[21]	0.0	51.8	75.6	56.5	72.5	76.9	75.5
GAMAN	0.0	4.0	6.6	15.0	22.2	16.7	9.9

ResNet

Perturb	∅	Proc	T	N	B	C	TNBC
None	91.1	91.1	90.6	72.0	87.2	89.3	90.8
BI[12]	0.0	10.9	36.8	24.8	32.8	45.3	26.3
GA	0.0	15.2	37.3	24.4	36.9	43.7	28.9
BI-S	0.0	13.0	43.4	27.4	35.8	51.5	29.9
GA-S	0.0	19.4	45.0	27.1	40.2	50.3	33.3
DF[21]	0.0	52.9	83.1	65.0	76.8	84.2	80.9
GAMAN	0.0	7.3	23.4	23.3	28.2	31.8	18.4

Table 9: Extended version of table 3 in the main paper for all four network architectures; additional AIP entries are denoted as gray cells. Robustness analysis of AIPs for various convnet architectures. AIPs are restricted to $||\cdot||_2 \leq 1000$. $(\mathsf{T}, \mathsf{N}, \mathsf{B}, \mathsf{C}) =$ (Translate, Noise, Blur, Crop).

AlexNet

User Θ^u	Recogniser Θ^r Proc	T	N	B	C	TNBC
GAMAN	1.4	6.4	9.2	13.5	12.3	5.6
/T	0.9	0.8	6.2	10.5	2.7	2.2
/N	1.2	4.2	4.8	11.7	9.5	3.9
/B	0.8	3.5	6.3	6.4	6.0	2.6
/C	2.4	2.5	9.2	13.1	1.3	3.4
/TNBC	0.6	1.2	4.5	7.8	2.9	1.9

VGG

User Θ^u	Recogniser Θ^r Proc	T	N	B	C	TNBC
GAMAN	1.6	2.1	8.5	11.8	5.6	3.5
/T	1.5	1.2	8.1	12.3	3.2	2.8
/N	2.0	2.5	3.9	12.6	6.7	3.9
/B	0.3	0.7	5.0	4.5	2.2	1.2
/C	2.0	1.6	9.5	14.0	1.9	3.1
/TNBC	0.6	0.7	4.3	7.3	2.3	1.4

GoogleNet

User Θ^u	Recogniser Θ^r Proc	T	N	B	C	TNBC
GAMAN	4.0	6.6	15.0	22.2	16.7	9.9
/T	2.5	2.3	11.6	18.5	7.2	4.9
/N	5.8	7.6	4.6	23.6	16.6	9.1
/B	0.4	0.8	8.6	5.8	3.1	1.4
/C	2.6	2.2	11.8	18.1	3.4	4.3
/TNBC	0.7	0.9	5.2	9.5	3.2	2.0

ResNet

User Θ^u	Recogniser Θ^r Proc	T	N	B	C	TNBC
GAMAN	7.3	23.4	23.3	28.2	31.8	18.4
/T	2.9	2.8	16.6	19.0	5.4	5.8
/N	5.3	12.9	4.2	23.5	20.1	10.2
/B	0.6	3.1	13.0	6.8	5.3	2.4
/C	3.5	3.1	17.0	18.8	3.2	5.4
/TNBC	0.7	1.2	6.5	9.3	2.9	2.3

Table 10: Extended version of table 4 in the main paper for all four network architectures. Recogniser's payoff table p_{ij}, $i \in \Theta^u$, $j \in \Theta^r$, for various convnet architectures. The user's payoff is given by $100 - p_{ij}$.

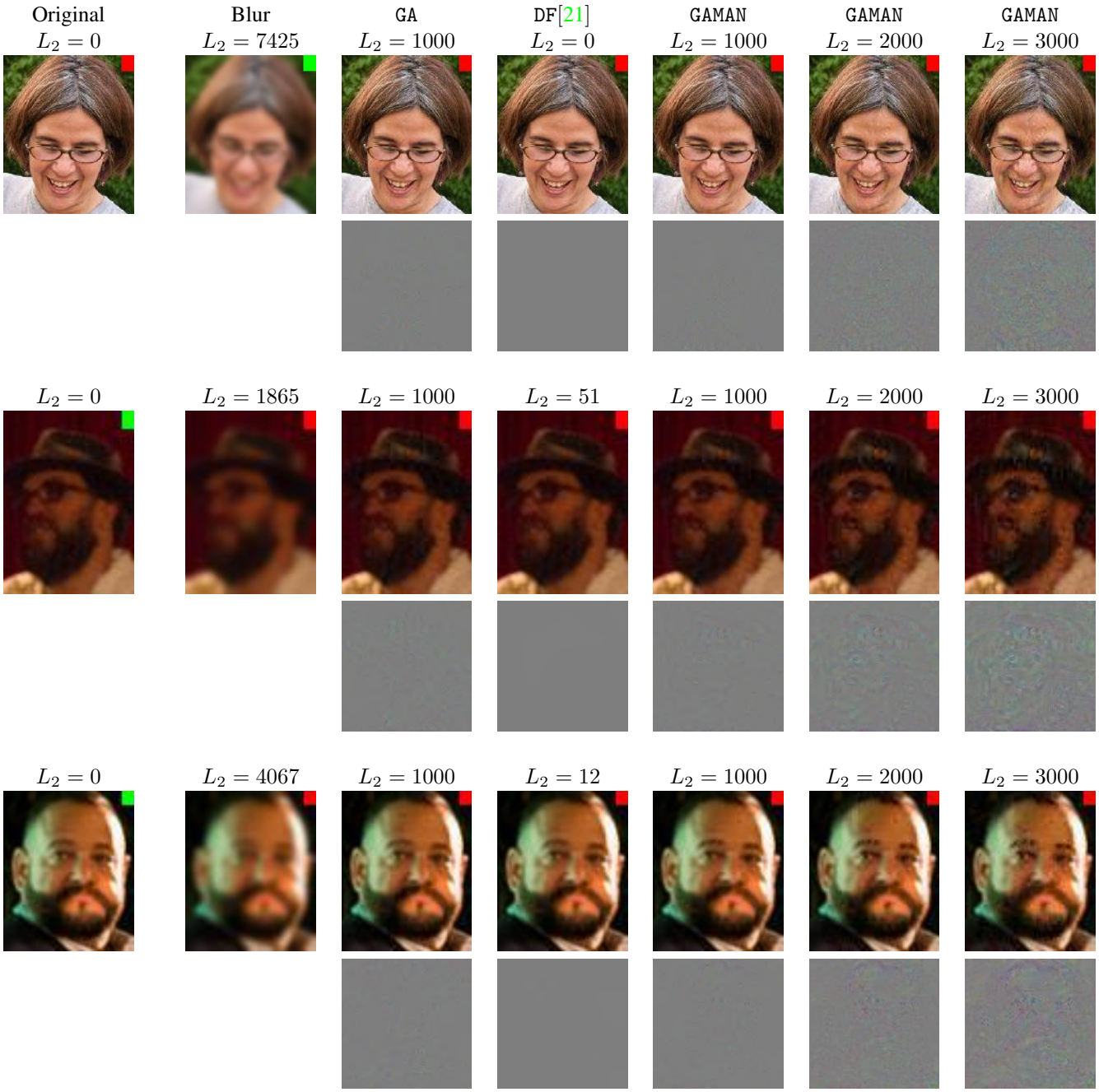

Figure 5: Randomly chosen perturbed images after Proc and the corresponding GoogleNet predictions (green for correct, red for wrong). Perturbations are visualised with gray background. GA and GAMAN reliably confuse the classifier at almost no cost on the aesthetics. As the L_2 norm increases, artifacts become more visible. Perturbations may be too small to be visible when printed; zoom in in electronic version for better visibility.

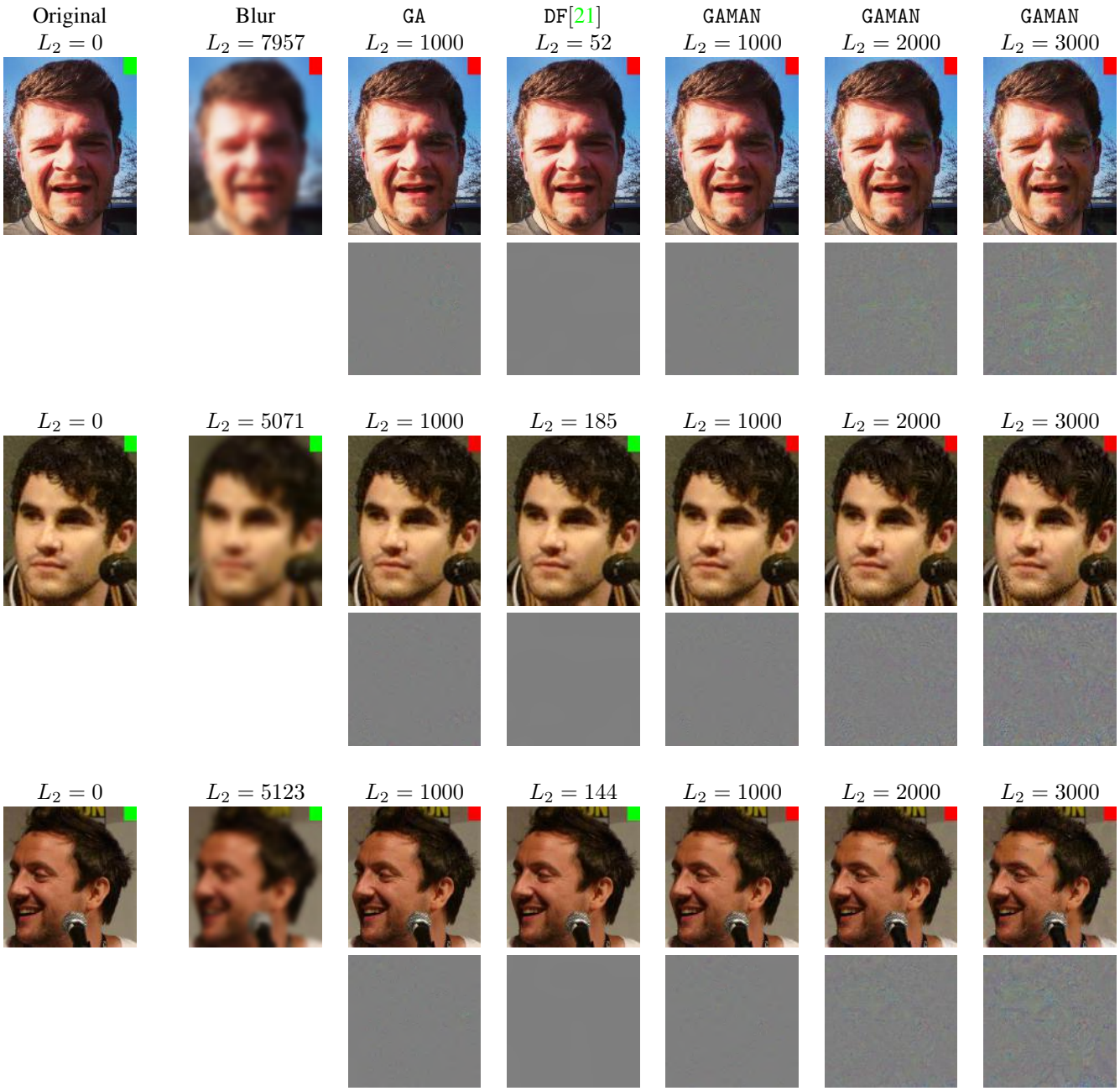

Figure 6: Randomly chosen perturbed images after Proc and the corresponding GoogleNet predictions (green for correct, red for wrong). Perturbations are visualised with gray background. GA and GAMAN reliably confuse the classifier at almost no cost on the aesthetics. As the L_2 norm increases, artifacts become more visible. Perturbations may be too small to be visible when printed; zoom in in electronic version for better visibility.

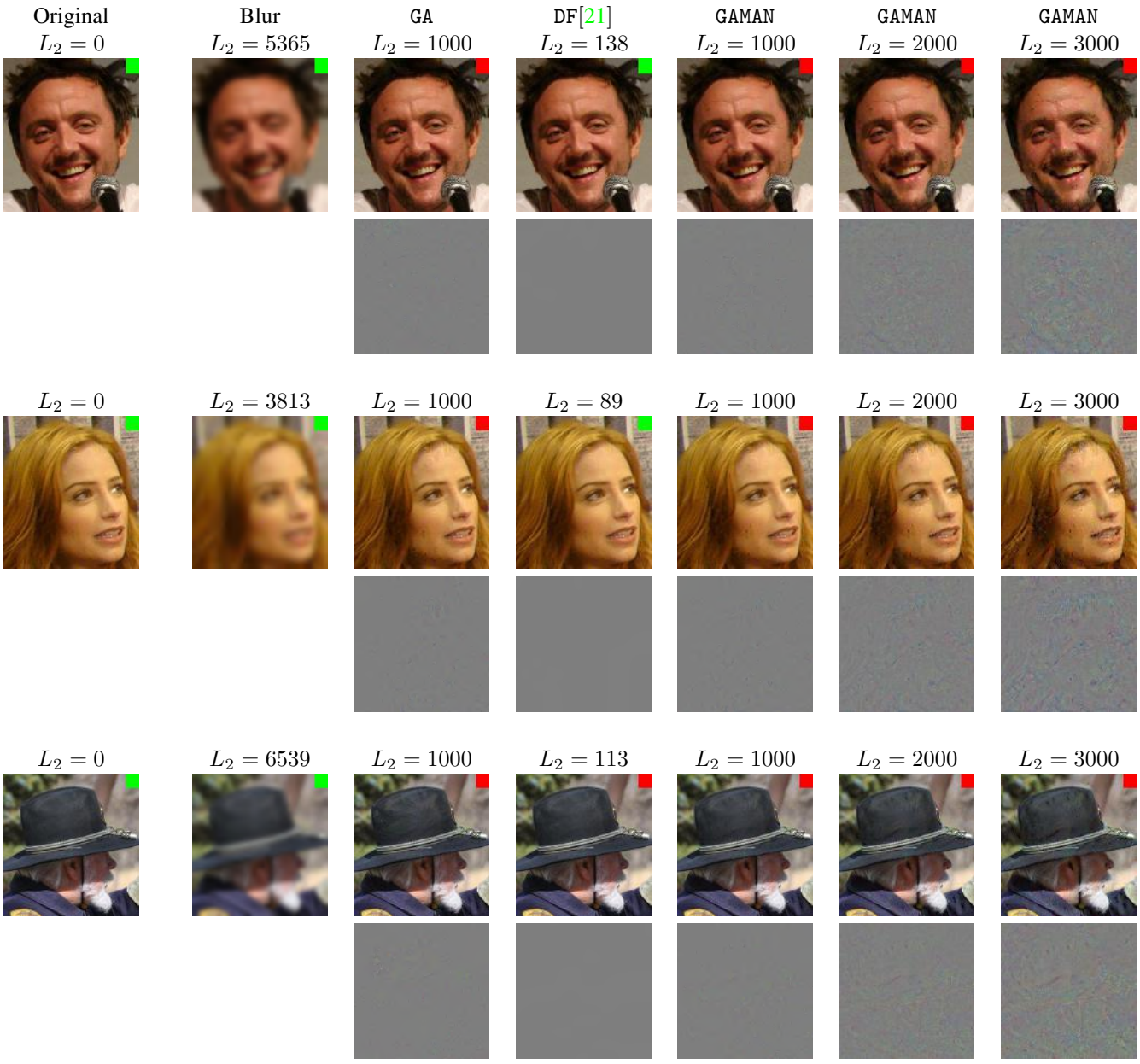

Figure 7: Randomly chosen perturbed images after Proc and the corresponding GoogleNet predictions (green for correct, red for wrong). Perturbations are visualised with gray background. GA and GAMAN reliably confuse the classifier at almost no cost on the aesthetics. As the L_2 norm increases, artifacts become more visible. Perturbations may be too small to be visible when printed; zoom in in electronic version for better visibility.

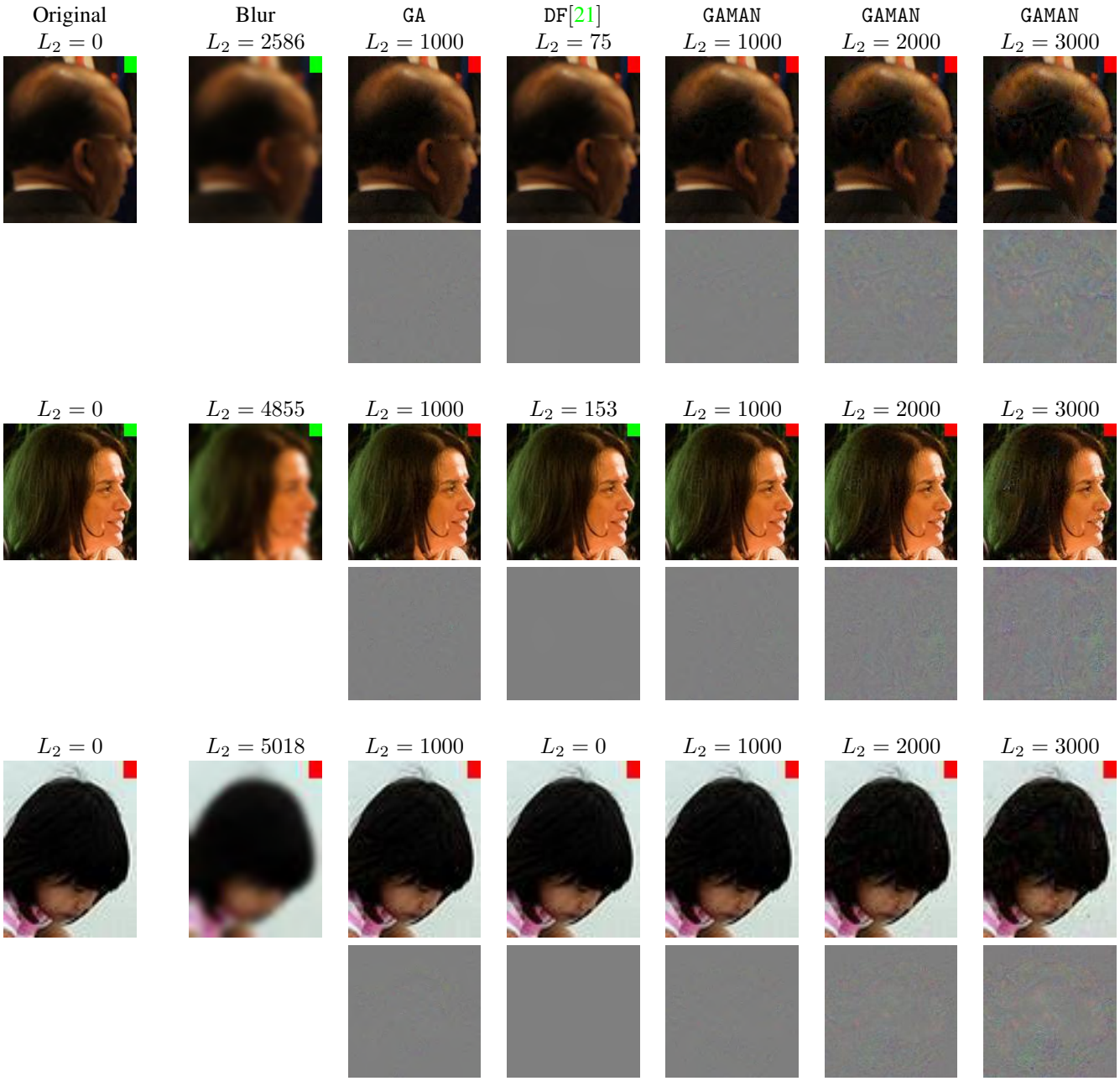

Figure 8: Randomly chosen perturbed images after Proc and the corresponding GoogleNet predictions (green for correct, red for wrong). Perturbations are visualised with gray background. GA and GAMAN reliably confuse the classifier at almost no cost on the aesthetics. As the L_2 norm increases, artifacts become more visible. Perturbations may be too small to be visible when printed; zoom in in electronic version for better visibility.