Natural and Effective Obfuscation by Head Inpainting

Qianru Sun[1]* Liqian Ma[2]* Seong Joon Oh[1]
Luc Van Gool[2,3] Bernt Schiele[1] Mario Fritz[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus
[2]KU-Leuven/PSI, Toyota Motor Europe (TRACE) [3]ETH Zurich
{qsun, joon, schiele, mfritz}@mpi-inf.mpg.de
{liqian.ma,luc.vangool}@esat.kuleuven.be vangool@vision.ee.ethz.ch

Abstract

As more and more personal photos are shared online, being able to obfuscate *identities in such photos is becoming a necessity for privacy protection. People have largely resorted to blacking out or blurring head regions, but they result in poor user experience while being surprisingly ineffective against state of the art person recognizers [17]. In this work, we propose a novel* head inpainting *obfuscation technique. Generating a realistic head inpainting in social media photos is challenging because subjects appear in diverse activities and head orientations. We thus split the task into two sub-tasks: (1) facial landmark generation from image context (e.g. body pose) for seamless hypothesis of sensible head pose, and (2) facial landmark conditioned head inpainting. We verify that our inpainting method generates realistic person images, while achieving superior obfuscation performance against automatic person recognizers.*

1. Introduction

Social media have brought about large-scale sharing of personal photos. While providing great user convenience, such a dissemination can pose privacy threats on users. It is essential to grant users an option to obfuscate themselves out of these photos. A good obfuscation method for social media photos should satisfy two criteria: *naturalness* and *effectiveness*. For example, putting a large black box over a person may be an effective obfuscation method, but would not be pleasant enough to share with friends.

Previous work on visual content obfuscation can be grouped into two categories: (1) *target-specific* and (2) *target-generic*. Some papers have proposed *target-specific* obfuscations, ones that are specialized against specific target machine systems, typically relying on adversarial examples [18, 23]. They yield nearly perfect identity protection with imperceptible changes on the input, but such a performance is guaranteed only against the targetted ones.

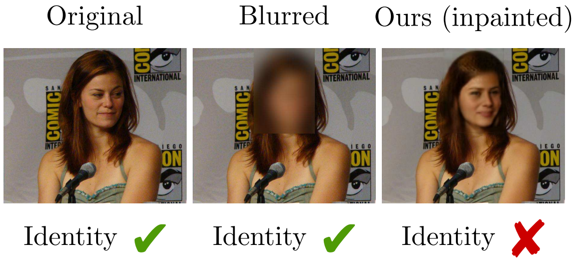

Figure 1: Our obfuscation method based on head inpainting generates much more natural patterns than common techniques like blurring, but still results in a more effective identity obfuscation against a recognizer.

On the other hand, *target-generic* obfuscations change the actual appearance of the person such that generic classifiers or even humans misjudge the identity. Commonly used obfuscation methods like black eye bar, face blurring, and blacking out head are examples of this type. These patterns, unfortunately, are neither visually pleasant nor effective against machine systems [17]. This paper proposes a *head inpainting* based approach to the target-generic identity obfuscation problem.

Generating realistic and seamless head inpainting on social media photos is hard. Subjects appear in diverse events and activities, resulting in varied backgrounds and head poses. Meanwhile, current generative face models are limited to frontal [3] or strictly aligned [12] faces.

We tackle the problem by factoring it into two stages. First, depending on the access to original face pixels, we either detect or generate facial landmarks. We detect them when we have access to the original face image, but when face has already been obfuscated, we generate (hypothe-

[*] Equal contribution.

size) them. The second scenario makes our approach versatile, by letting us e.g. *upgrade* existing weak obfuscations on the web including blacked out or blurred out heads (called blackhead and blurhead in the remainder of the paper) into our novel privacy-enhanced versions. Then, conditioned on the face landmarks, we inpaint a realistic head that blends naturally into the context. We show that the resulting head-inpainted images mislead machine recognizers.

Our key contributions are: (1) Novel natural, effective obfuscation methods based on head inpainting; (2) Novel landmark guided image generation approach for both head visible and blackhead cases in challenging social media photos; (3) Novel facial landmark generator that effectively hypothesize realistic facial structures and poses given context (blackhead scenario).

2. Related work

Identity obfuscation. A few works from the vision community have analyzed and developed obfuscation patterns for avoiding person identification. First, we introduce a line of work on *target-generic* obfuscations that are designed to work against generic automatic person recognizers as well as humans. Oh *et al.* [17] and McPherson *et al.* [15] have analyzed the obfuscation performance of blacking or blurring faces against automatic recognizers. They have concluded that these common obfuscation methods are not only unpleasant but also ineffective, in particular due to the adaptability of convnet-based recognizers [17]. More sophisticated approaches have been proposed since then. Hassan *et al.* [8] have proposed to mask private image content via *cartooning*. Brkic *et al.* [1] have generated full-person patches to overlay on top of person masks. Similarly, we propose an obfuscation technique based on *head inpainting*. The key difference is that while [1] generates persons with uniform poses independent of the context (fashion photos), we naturally blend generated heads with diverse poses into varied background and body poses (social media photos).

For the *target-specific* obfuscations, Oh *et al.* [18] and Sharif *et al.* [23] have proposed *adversarial example* based obfuscation. While the obfuscation performance is superb even at imperceptible perturbation level, such a performance depends highly upon the accessibility to target system's inner parameters. Since we aim to obfuscate identities against a wide range of recognition systems, we do not condition our inpainting against a specific recognizer.

Image inpainting. In our work, we propose generative adversarial network (GAN) based method to complete head regions based on the context. Raymond *et al.* [31] and Pathak *et al.* [19] have also used GANs to inpaint pixels based on the context. However, both approaches assume appearance and texture similarity between the missing part and the context. Our approach can inpaint heads solely from body and scene context, without resorting to any information from the head region. In particular, while [31] inpaints aligned faces, we inpaint heads in the challenging social media setup in which people appear with diverse poses and backgrounds by taking a two-stage approach.

Structure guided image generation. For generating realistic head inpainting that naturally blends into the given body pose and scene context, we have conditioned the inpainting on face landmarks. Prior work on structure-guided image generation has shown that such a guidance is indeed very helpful for generating images with complex inner structures (e.g. persons) [13, 5, 28, 30, 6, 33, 14, 2]. Ma *et al.* [13] have trained a system to synthesize persons based on pose. Similarly, Walker *et al.* [28] have used the predicted future poses to condition a GAN to generate future frames in videos. In [30], Wang and Gupta factorizes the indoor scene generation task into surface normal generation and texture imbuing stages. Ehsani *et al.* [6] addresses the object occlusion problem by first predicting the contour of the invisible parts and then generating the appearance inside the contour. Alpher *et al.* [5] have generated faces conditioning on detected face landmarks. Despite the similarity shared by our work, [5] only generates well-aligned faces. Our approach generates realistic, seamless head patches in social media photos where the body pose and the background are very diverse.

3. Head inpainting framework

We focus on the scenario where the user wants to obfuscate some identities in a social media photo by inpainting new heads for them. The task is challenging due to complex poses and background typical in social media photos. We use facial landmarks to provide strong guidance for the head inpainter. We factor the head inpainting task into two stages: (1) landmark detection or generation and (2) head inpainting conditioned on body context and landmarks.

Figure 2 describes the global view of our two-stage approach. It takes either the original or blackhead image[1] as input, in order to give flexibility to deal with cases where the original images are not available. Given original or head-obfuscated input, stage-I detects or generates landmarks, respectively. Stage-II takes the blackhead image and landmarks as input, and outputs the generated image.

3.1. Stage-I: Landmark

In stage-I, we detect or generate face landmarks to guide head inpainting in the subsequent stage. An overview of stage-I is shown in Figure 3. For landmark detection, we detect 68 facial keypoints using the python dlib toolbox [10]. For landmark generation, we train the Landmark Generator (G_L) adversarially with the Discriminator (D_L). We will

[1] Blurhead image is another important obfuscation, and it is easily adapted in our approach. We use blackhead image as a default example.

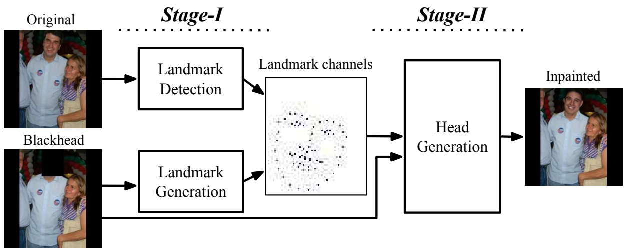

Figure 2: Our two-stage head inpainting framework. The input of stage-I is either the original or the blackhead image. The output is the inpainted image.

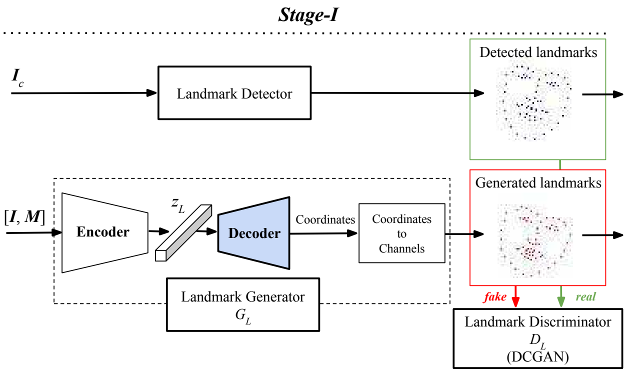

Figure 3: Stage-I: Landmark Detection/Generation. The detector takes the original image $\boldsymbol{I_c}$ as input; the generator takes a blackhead image $\boldsymbol{I}$ and the head mask $\boldsymbol{M}$ as input. For the Decoder in G_L, we consider three versions of training: from scratch (Scratch); pre-training as autoencoder (AEDec); pre-training as Point Distribution Model (PDMDec).

describe the landmark generator in greater detail.

Landmark Generator (G_L). G_L has an autoencoder structure with two parts: Encoder and Decoder. The Encoder compresses the body/scene context of the blackhead image to a latent vector. The Decoder then decodes the vector into landmark coordinates. In the following, we describe details of the Encoder and Decoder.

Encoder of G_L. Encoder takes a blackhead image $\boldsymbol{I}$ and the corresponding head mask $\boldsymbol{M}$ (indicating the head bounding box) as inputs. Encoder maps the input $\boldsymbol{X} = [\boldsymbol{I}; \boldsymbol{M}]$ to a latent vector $\boldsymbol{z}_L$. Encoder has 6 convolutional residual blocks; the latent vector $\boldsymbol{z}_L$ is 32-dimensional.

Decoder of G_L. Taking the latent vector $\boldsymbol{z}_L$ as input, Decoder generates 2×68 landmark coordinates $\boldsymbol{L}$. Decoder contains 6 fully connected residual blocks. Both Encoder and Decoder are trained from scratch by default.

Training Encoder and Decoder from scratch is challenging due to diverse body pose and background clutter in social media photos. Therefore, we consider first training a strong decoder and training the encoder from scratch with respect to the trained (and fixed) decoder. Such a procedure is inspired by the previous work on knowledge transfer between deep models trained on different tasks [7, 22].

We consider pre-training Decoder in three possible ways: (1) from scratch (and simultaneously training with Encoder), (2) autoencoder, and (3) using the Point Distribution Model (PDM, [4]).

AE decoder (AEDec). The autoencoder reconstructs face landmarks using an encoder and a decoder through a bottleneck layer. Both are fully connected layers with ReLU activations. L_2 loss is as the loss function.

PDM decoder (PDMDec). We consider using the Point Distribution Model (PDM) to better represent the 3D pose variations [4, 32][2]. We train the PDM over the detected landmarks on PIPA *train* set images. Our landmark points are parametrized using $\boldsymbol{p} = [s, R, t, q]$ denoting scale, orientation, translation and non-rigid transformations, respectively. The PDM decoder has the following formulation:

$$\boldsymbol{L} = s \cdot \boldsymbol{R} \cdot (\bar{\boldsymbol{L}}_{3D} + \boldsymbol{\Phi q}) + \boldsymbol{t} \tag{1}$$

where $\bar{\boldsymbol{L}}_{3D}$ denotes the mean value of the 3D landmarks mapped from our 2D data, and $\boldsymbol{\Phi}$ the $3 \times n$ principal component matrix. The output $\boldsymbol{L}$ has $n + 6$ parameters. In the experiments we use $n = 34$ principal components.

Loss functions of G_L and D_L. We use the L_2 loss as well as an adversarial loss for optimization. Landmarks trained only with the L_2 loss show noisy alignments; we found the adversarial loss to be useful at remedying this. We adopt the DCGAN discriminator [20]. The landmark coordinates are converted to channels to input to the convolutional layers, where the conversion process is differentiable. We have also tried a fully-connected discriminator, instead of the DCGAN discriminator, but the difference was marginal.

For training D_L, any landmark generated by G_L are labeled *fake*, while we use the *detected* landmarks as the *real* examples. Exact losses are formulated as follows:

$$\begin{aligned}\mathcal{L}_{D_L} =& \mathbb{E}_{\boldsymbol{X} \sim p_{data}(\boldsymbol{X})} \big[\log D_L(\boldsymbol{X})\big] + \\ & \mathbb{E}_{\boldsymbol{X} \sim p_{data}(\boldsymbol{X})} \big[\log\left(1 - D_L(G_L(\boldsymbol{X}))\right)\big], \end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{L}_{G_L} =& \mathbb{E}_{\boldsymbol{X} \sim p_{data}(\boldsymbol{X})} \big[\log\left(D_L(G_L(\boldsymbol{X}))\right)\big] + \\ & \lambda_L \|G_L(\boldsymbol{X}) - \boldsymbol{L}_d\|_2, \end{aligned} \tag{3}$$

where $\boldsymbol{X}$ is the concatenation of the obfuscated image $\boldsymbol{I}$ (3 channels) and the head mask $\boldsymbol{M}$ (1 channel). $\boldsymbol{L}_d$ is the detected landmark coordinates (*ground truth*). $\lambda_L \geq 0$ is a scalar weight.

[2] We use [32] to train the PDM model [4]. Non-rigid structure from motion [27] is used to map 2D points to 3D in this code. Our training data are the detected landmarks in PIPA TRAIN set.

3.2. Stage-II: Inpainting

Stage-II generates the head inpainting based on the landmarks from Stage-I and the blackhead or blurhead image. Figure 4 shows an overview; the head generator G_H is trained adversarially with a head discriminator D_H.

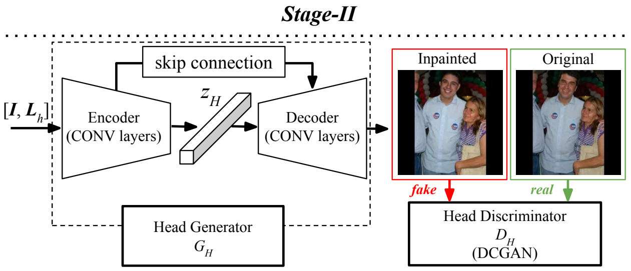

Figure 4: Stage-II: Head generation. The input are blackhead image $\boldsymbol{I}$ and landmark channels $\boldsymbol{L}_h$. The generator has an Auto-encoder structure which encodes the input to a bottleneck then decodes to a fake image. The discriminator is the same as in DCGAN [20].

Input. The 68-channel landmark heatmaps $\boldsymbol{L}_h$ from Stage-I are concatenated with the blackhead (or blurhead) image $\boldsymbol{I}$ as an input to the generator G_H. The landmark heatmaps provide the missing skeleton information in the obfuscated image.

We treat the blackhead image as *fake* and the original image as *real* the head discriminator D_H. Note that we use the whole body image instead of just head regions to provide sufficient information about the body and background to generate a realistic inpainting.

Head Generator (G_H) and Discriminator (D_H). The head generator G_H has a a convolutional autoencoder with skip connections between encoder and decoder, inspired by the U-Net [21]. The skip connections propagate image information directly from input to output, improving the fine-grained details in the output. The architecture of the head discriminator D_H is the DCGAN discriminator [20].

Loss function. We use the L1 and the adversarial losses to optimize G_H and D_H:

$$\begin{aligned}\mathcal{L}_{D_H} =& \mathbb{E}_{\boldsymbol{Y}\sim p_{data}(\boldsymbol{Y})}\left[\log D_H(\boldsymbol{Y})\right]+\\ & \mathbb{E}_{\boldsymbol{Y}\sim p_{data}(\boldsymbol{Y})}\left[\log\left(1-D_H(G_H(\boldsymbol{Y}))\right)\right], \qquad (4)\\ \mathcal{L}_{G_H} =& \mathbb{E}_{\boldsymbol{Y}\sim p_{data}(\boldsymbol{Y})}\left[\log\left(D_H(G_H(\boldsymbol{Y}))\right)\right]+\\ & \lambda_H\|G_H(\boldsymbol{Y})-\boldsymbol{I}_c\|_1, \qquad (5)\end{aligned}$$

where $\boldsymbol{Y}$ is the concatenation of the obfuscated image $\boldsymbol{I}$ and the landmark heatmaps $\boldsymbol{L}_h$. $\boldsymbol{I}_c$ is the original image. $\lambda_H \geq 0$ is a scalar weight. Detailed architecture and hyperparameters are given in the supplementary materials.

4. Experiments

We evaluate the presented two-stage head inpainting pipeline on a social media dataset in terms of inpainting appearance and pose plausibility, as well as the identity obfuscation performance against machine recognizers. We analyze the impact of different input types (original, blackhead, and blurhead), different choices of landmark decoders, and the losses for the landmark generators (§3.1).

4.1. Dataset

We use the PIPA dataset [34], the largest social media dataset to date with people in diverse events, activities, and poses. It is a suitable for evaluating our methods under the social media obfuscation scenario.

In order to maximize the amount of training data, we have introduced a new partitioning of the images in PIPA. We partition 2,356 PIPA identities into `TRAIN` set (2,099 identities, 46,576 instances) and `TEST` set (257 identities, 5,175 instances). We have further pruned both partitions with heavy profile or back-view heads, resulting in 34,383 instances in `TRAIN` and 1,909 in `TEST`. The `TRAIN` set is used for training landmark and head generators. `TEST` set is the evaluation set.

Our landmark and inpainting generators take a fixed-size image ($256 \times 256 \times 3$) as input. For every training and testing sample, we prepare the input by first obtaining the *body crop*, following the procedure in [16, 25]: extend the head box with fixed ratios (3×width and 6×height), and then resize and zero-pad the body crop such that it fits tightly in the square 256×256.

4.2. Scenarios and inputs

Our approach introduced in §3 is versatile and supports scenarios where the user (who wants to obfuscate an image) has access to the original image or only has access to already head-obfuscated images (e.g. blacked out). The necessity for this versatility is that social network service providers may aim to upgrade the privacy level by obfuscating images through blurring or blacking-out heads, even though it has been shown to be quite ineffective [17].

In order to simulate multiple scenarios, we consider three types of inputs to our obfuscator: original, blackhead, or blurhead, where the latter two are common obfuscation techniques these days. We prepare blackhead and blurhead inputs following the procedure in [17]. PIPA head box annotations indicate the head region to be obfuscated, which is either filled in with black pixels or smoothed with a Gaussian blur kernel specified in [17].

4.3. Quantitative results

Our head inpainting should both look natural and effectively obfuscate the identity. We report quantifiable measurements of the two criteria in this section.

Table 1: Evaluation of proposed obfuscation methods. We quantify the quality of the proposed obfuscation method against landmark quality, inpainting quality, as well as obfuscation effectiveness (person recognition rates). We vary the loss (D_L here represents the adversarial loss) and decoder used in our landmark generator (§3.1); the head inpainter is always the G_H + D_H(§3.2).

Obfuscation method			Evaluation							
	Landmark		Landmark		Inpainting		Person recognizer		
Input	Loss	Decoder	L_2	Norm. L_2	SSIM	mask-SSIM	head	body+head	head contrib.
Original	No head inpainting		/	/	1.000	1.000	85.6%	88.3%	72.2%
Original	NN head copy-paste		/	/	0.872	0.195	1.2%	7.1%	67.5%
Blur	No head inpainting		/	/	0.931	0.396	52.2%	71.6%	3.2%
Blur	Detected landmarks		0.00	0.000	0.962	0.679	43.7%	51.7%	70.8%
Blur	L_2	Scratch	6.32	0.230	0.954	0.578	36.2%	48.4%	66.8%
Blur	L_2+D_L	Scratch	4.85	0.182	0.955	0.586	38.0%	48.4%	66.6%
Blur	L_2+D_L	AEDec	4.77	0.180	0.951	0.585	37.5%	48.0%	66.1%
Blur	L_2+D_L	PDMDec	4.50	0.168	0.953	0.593	37.9%	49.1%	66.7%
Black	No head inpainting		/	/	0.815	0.000	2.1%	67.0%	14.0%
Black	Detected landmarks		0.00	0.000	0.902	0.405	10.1%	21.4%	70.8%
Black	NN landmarks		2.48	0.088	0.896	0.332	7.9%	20.4%	71.3%
Black	L_2	Scratch	13.6	0.501	0.884	0.186	5.8%	17.4%	73.6%
Black	L_2+D_L	Scratch	13.0	0.477	0.882	0.191	5.8%	17.2%	71.4%
Black	L_2+D_L	AEDec	11.7	0.431	0.885	0.199	5.6%	17.4%	72.5%
Black	L_2+D_L	PDMDec	12.3	0.453	0.885	0.196	5.6%	17.4%	71.0%

4.3.1 Landmark

As intermediate output, facial landmarks should represent a realistic human face in order to provide correct guidance to the inpainting stage. In this section, we evaluate the generated landmark quality in terms of the L_2 distance to the detected landmarks, assuming that the detected landmarks are accurate. The L_2 distances are normalized with respect to the inter-ocular distances [10].

We investigate three axes of factors for our landmark generator. (1) Input type: original, blackhead, or blurhead. (2) Loss function: only L_2 versus L_2 and adversarial loss (D_L). (3) Decoder type: trained from scratch, autoencoder pretrained (AEDec), or Point Distribution Model pretrained (PDMDec). A summary of the results is in Table 1 ("Landmark" column). On the original images, our best landmark generator achieves the L_2 distance of 2.41 on average (not shown in table), which gives an upper bound on the landmarks generated on blurhead or blackhead.

Input type. We compare the L_2 distance between the generated and detected landmarks for three types of inputs: original, blackhead, or blurhead. For original images, we use detected landmarks (by definition zero L_2 distance). We observe from Table 1 that blurhead inputs show more accurate landmarks than blackhead cases: e.g. 6.32 (blur) versus 13.6 (black) for the baseline landmark generator (L_2 loss, trained from scratch). Blurhead images already provide much structural information.

Loss function. We compare two loss functions: without adversarial loss (L_2) versus with adversarial loss ($L_2 + D_L$). Given a blackhead input with landmark decoder trained from scratch, using only L_2 loss yields the 13.6 distance, while adding D_L marginally improves the distance to 13.0. However, for blurhead images, the improvement due to D_L is much greater (from 6.32 to 4.85).

Decoder. We consider three choices of decoder in the landmark generator G_L: learning from scratch (Scratch), pre-trained with AE (AEDec), and pre-trained with PDM (PDMDec). For both blurhead or blackhead cases, conditioning the decoder with either AEDec or PDMDec helps generating better landmarks: e.g. for blackhead input, L_2 distance improved from 13.0 to 11.7 and 12.3, respectively.

4.3.2 Inpainting

Head inpainting, the final output of our method, should look natural to be suitable as a social media photo. While we will visualize the output and report user study in the next sections (§4.4 and §4.5), we provide a large-scale summary measures for the quality of final output using the SSIM distance [29] from the original image. We report two measures: comparing the whole images (SSIM [29]) and head region only (mask-SSIM [13]). As a baseline, we consider inpainting with the Nearest Neighbor (NN) head[3]. Our head

[3] NN head is searched in training data based on the mean L_2 distance of detected landmarks.

inpainting always gives a better SSIM (>0.88) than this baseline (0.872).

4.3.3 Obfuscation

We now measure how well our inpainting obfuscates corresponding identities. Unlike some prior work [23, 18], our obfuscation scheme is *target generic* – it is designed to actually change the identity, instead of fooling specific classifiers. We use state of the art person recognizers [16, 18] to measure the change in identifiability due to obfuscation. While the method is target generic, we report results on two recognisers with great structural differences to provide further evidence that the obfuscation is effective regardless of the target (experiments on one of the recognisers are in supplementary materials). We provide a rationale for our good obfuscation performance based on the analysis of the recognizer attention. We show furthermore that the obfuscation results in non-confident top-1 predictions – the obfuscation does not change the appearance to another person known to the recogniser (which may be unethical) but comes up with a new, unseen identity.

Person recognizer. We use the social media person recognition framework `naeil` [16]. Unlike typical face recognizers, `naeil` uses body and scene context cues for recognition. It has thus proved to be relatively immune to common obfuscation techniques like blacking or blurring head regions [17].

Following [16], we first train feature extractors over head and body regions, and then train an SVM identity classifier on top of those features. We may also concatenate features from multiple regions (e.g. head+body) to allow it to extract cues from multiple regions. In our work, we use GoogleNet [26] features from `head` and `head+body` to evaluate obfuscation performances. AlexNet [11] based recogniser is also considered in Supplementary Materials to show that the obfuscation is similarly effective for two greatly distinct types of recognisers (e.g. 98 layers for GoogleNet and 8 layers for AlexNet).

Head inpainting provides good protection. Table 1 shows obfuscation performance (columns `head` and `head+body`). Under no obfuscation, the `head+body` recognition performance is 88.3%. Black/blurring baselines give 67.0%, and 71.6%, respectively – confirming the observation in [17] that these are ineffective. On the other hand, our head inpainting methods show $< 50\%$ (blurhead input) and $< 21\%$ (blackhead input) recognition rates for `head+body` recognizers. They are more effective protection techniques than blacking or blurring head regions.

Cues used. We compare the recognition rates between `head` and `head+body`. When the recognizer relies solely on head cues, while the head has been inpainted, then the recognition rates are lower than the `head+body` counterparts. For example, the last row method against `head` recognizer gives 5.6% versus 17.4% for `head+body`, nearly reaching the chance level recognition rate 2.1%.

Input type. While having access to blurred head images help generating more plausible landmarks (§4.3.1) as well as visually natural head inpainting (§4.4), they may leak identity information. We compare the recognition rates when either blurhead or blackhead inputs are used. Our head inpainting based on blackhead result in $17\% \sim 21\%$ accuracy, while blurhead based results are in the range $48\% \sim 50\%$ accuracy. The choice of input type gives users a control over the trade-off between plausibility of generated heads and the obfuscation performance.

Detected versus generated landmarks. While identity information may leak through blurred heads, it may also leak through the landmark detections (face shape). On the other hand, generated landmarks enjoys the possibility to come up with an equally plausible landmark hypothesis but with different face shapes. For the blackhead input, the detected landmarks indeed result in higher recognition rate (21.4%) than generated ones (e.g. 17.4% on last row), with similar trend for the blurhead cases.

Rationale for good obfuscation – recognizer attention. We have verified that our head obfuscation scheme exhibits better performance than commonly used ones like blacking and blurring. We give a rationale for this phenomenon by means of the *recognizer attention*. Given an input, *recognizer attention* refers to the image regions where recognizers extract cues from. We hypothesize that while blacked or blurred heads induce recognizer attention on non-head regions, our inpainted heads attract attention on the heads.

For the *recognizer attention* we have used the gradient-based mechanism from Simonyan *et al.* [24]. We first compute the gradient of the neural network prediction with respect to the input image; take maximal absolute values along the RGB channel; and then smooth with Gaussian blurring. To quantify the chance of attending on the head region, we have computed the "head contribution" score by estimating head contrib. $= \mathbb{P}[\text{max attention is inside head region}]$ over the test samples.

See final column of table 1 for the results. We observe that while the original image has 72.2% chance of inducing attention on the head region, blacked or blurred heads are much less likely to attract the recognizer's attention (14.0% and 3.2%, respectively). This explains why `head+body` is still performing well: it simply ignores the confusing head cue. On the other hand, our inpainting-based obfuscation still attracts the recognizer's attention as much as the non-obfuscated head image does (71.0% versus 72.2%). This indicates that the realism of inpainted heads encourages the recognizer to still rely its decision on the inpainted head, effectively leading to misjudgment by the recognizer.

Low prediction confidence and ethics. Ethical problems might entail if the obfuscation mislead the recogniser into confidently predicting other identities in the gallery set. We have measured the SVM prediction confidence (1-vs-all SVM) on the original as well as obfuscated images to ensure that the obfuscation results in a uniformly low prediction scores.

On the original images, the top-1 identity is predicted with SVM score 0.63 on average. On the other hand, our inpainting conditioned on blurhead results in -0.29 average top-1 SVM score, inpainting conditioned on blackhead results in much lower top-1 score of -0.52. This confirms that the inpainting based obfuscation does not shift the identity prediction to another person with high confidence. If the recogniser filters out low-confidence predictions, a common practice in application, then the head-inpainted images will most likely be filtered out as "background identity".

4.4. Qualitative results

For confirming the naturalness of the inpainted heads, we have measured the SSIM score in §4.3.2. However, SSIM is only a proxy measure. In this section, we qualitatively visualize the quality of the generated landmarks as well as inpainted heads. We also include user study in the next section (§4.5)

For generating natural heads, landmarks should look like that of an actual face and be consistent with the body pose. However, at the same time obfuscation performance benefits from landmarks that do not preserve the original face shape. In this section, we discuss if our generated landmarks achieve both realism, while effectively obfuscating machine recognizers. Qualitative results are in Figure 5.

Detected versus generated landmarks. Given an original image with a visible head, we detect landmarks, while for blackhead we hypothesize them from regions other than the head itself. The comparison between columns 2,3 (detected landmarks) and columns 4,5 (generated landmarks) in Figure 5 illustrates the difference. In all the examples shown, the detected landmarks closely follow the original image. On the other hand, the generated landmarks, especially for blackhead cases, results in landmarks and head inpainting with different head poses. However, the generated landmarks are still plausible with respect to the body pose and activity. Finally, note that by generating landmarks, we can further mask identity information (recognition rates are consistently lower for inpainting based on generated landmarks), while keeping reasonable realism.

Blackhead versus blurhead. Landmarks may be generated from either blurhead or blackhead images. We visualize how the head information contained in blurred cases improve the inpainting quality. Columns 2,4 and columns 3,5 in Figure 5 show respective examples for blur and black cases. Involving blurred head images during landmark and head generation results in inpainting that resembles the original head, especially the head pose and hair color/style (*e.g.* ID-690). On the other hand, not providing any information in the head region results in a significantly different, yet plausible, head images. In particular, when even landmarks are generated, the resulting head images are drastically different from the original one. Such a shift of appearance is reflected in the low recognition rate (17.4%).

Table 2: Human perceptual study (HPS) scores and landmark detection success ratios (LDSR). Landmarks are from "detected", and "generated" by PDMDec methods.

			blurhead(Ours)		blackhead(Ours)	
	Orig.	CE [19]	detected	generated	detected	generated
HPS:	0.93	0.04	0.60	0.39	0.19	0.11
LDSR:	1.00	0.36	1.00	0.95	0.99	1.00

4.5. Comparing against the state-of-the-art

In this section, we compare the quality of our inpainting against two state-of-the-art inpainting methods [19, 1] via an extensive user study. We did not compare directly against [1] because it focuses on full body replacement using body contours and the generated heads are visually far from being competitive (e.g. Figure 1 in [1]). For all methods, we perform the human perceptual study (HPS) on Amazon Mechanical Turk (AMT). For each method, we show 55 real and 55 inpainted images in a random order to 20 users. Users press the real or fake button for an image within a second. The first 10 images are only practice samples [13, 9].

Table 2 shows comparison of the considered methods. The first row contains the ratios of images that were judged as real for different methods: (1) original unaltered; (2) inpainted by the Context Encoder (CE) [19] (blackhead image as input); (3) inpainted by our four models. We observe several interesting results. (1) Only 93% of the participants believed the original image to be real; this gives an upper bound on the score. (2) Our method based on blackhead images with generated landmarks results in 11% of the users believing that the image is real – nearly threefold boost from the CE baseline (4%). (3) Conditioning on the blurhead helps a lot (from 11% to 39% for generated landmarks) (4) Detected landmarks greatly improve the realism compared to generated ones (from 39% to 60% for blurhead). The realism is not perfect yet, but we greatly outperform the prior state of the art.

Finally, we also measure the landmark detection success ratio (LDSR) as a proxy measure of the output soundness (inspired by [9]). Intuitively, LDSR should be higher for heads with greater realism. As shown in Table 2, heads inpainted by our methods have LDSR above 95%, while ones inpainted by CE achieve only 36%. Our methods generate heads with much clearer face structures.

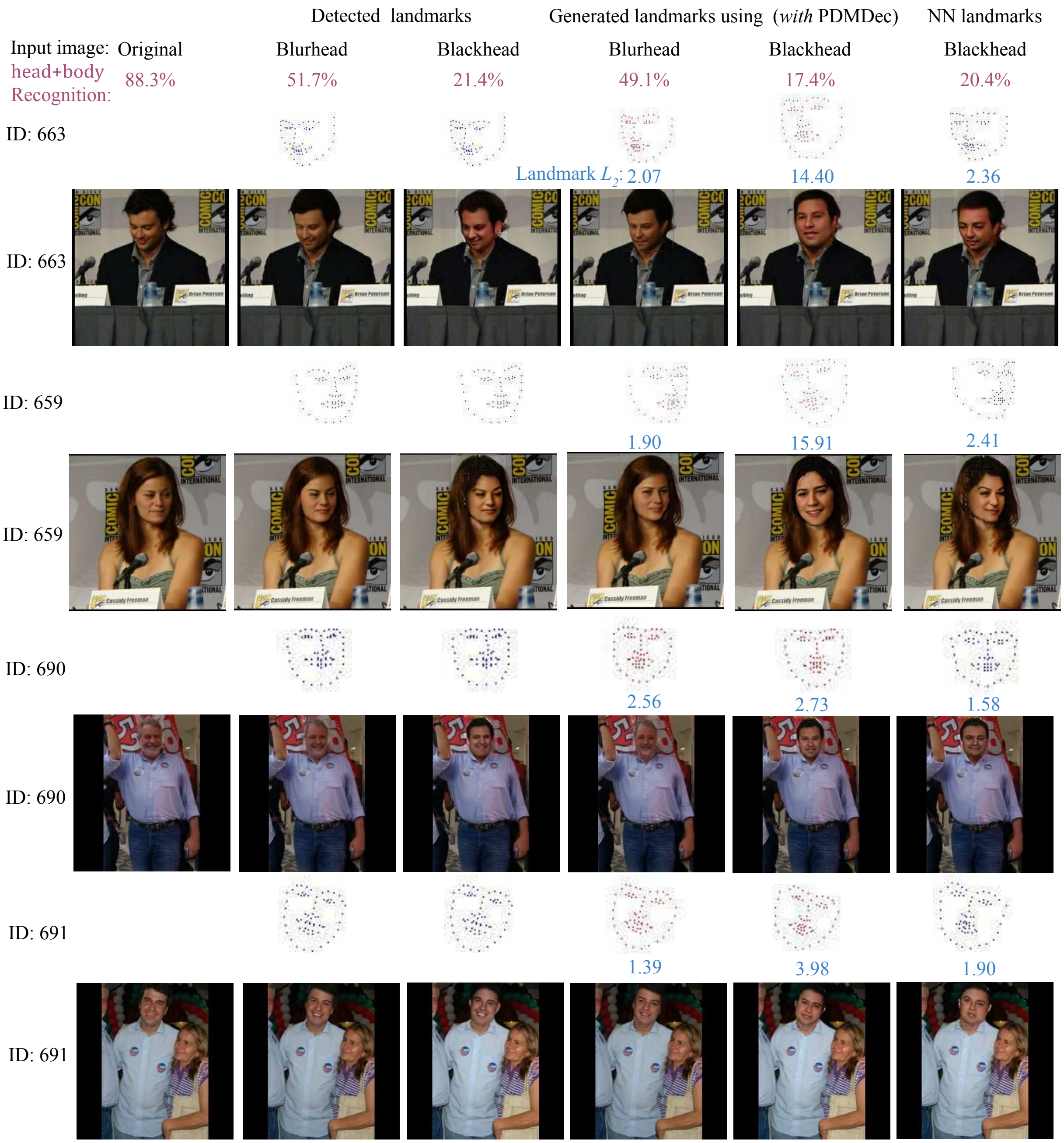

Figure 5: Head inpainting results using detected and generated landmarks (from the PDMDec model). Top rows present key quantitative numbers. The L_2 distance between detected and generated landmarks is also given for each single instance.

5. Conclusion

To address the problem of obfuscating identities in social media photos, we have presented a two-stage head inpainting method. Despite the challenges in the social media setup (diverse head and body poses and backgrounds), our method has proved to generate both natural obfuscation patterns that effectively confuses automatic person recognizers. In particular, our method is *target-generic*: the obfuscation is not conditioned on a particular recognizer, be it human or machine. Also, the method does not require access to the original image, enabling to "upgrade" weak obfuscation patterns (e.g. blurred or blacked heads) to our privacy-enhanced version.

Acknowledgments

This research was supported in part by German Research Foundation (DFG CRC 1223), Toyota Motors Europe.

References

[1] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1319–1328, July 2017. 2, 7

[2] L. Chen, H. Zhang, J. Xiao, W. Liu, and S. Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[3] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3395, 2017. 1

[4] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 3

[5] X. Di, V. A. Sindagi, and V. M. Patel. GP-GAN: Gender preserving gan for synthesizing faces from landmarks. *arXiv*, 1710.00962, 2017. 2

[6] K. Ehsani, R. Mottaghi, and A. Farhadi. SeGAN: Segmenting and generating the invisible. *arXiv*, 1703.10239, 2017. 2

[7] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2827–2836, 2016. 3

[8] E. T. Hassan, R. Hasan, P. Shaffer, D. J. Crandall, and A. Kapadia. Cartooning for enhanced privacy in lifelogging and streaming videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1333–1342, 2017. 2

[9] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2017. 7

[10] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 2, 5

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 6

[12] Z. Lu, Z. Li, J. Cao, R. He, and Z. Sun. Recent progress of face image synthesis. *arXiv*, 1706.04717, 2017. 1

[13] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation. In *The 31st Annual Conference on Neural Information Processing*, pages 405–415, 2017. 2, 5, 7

[14] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[15] R. McPherson, R. Shokri, and V. Shmatikov. Defeating image obfuscation with deep learning. *arXiv*, 1609.00408, 2016. 2

[16] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *International Conference on Computer Vision*, 2015. 4, 6

[17] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition; privacy implications in social media. In *European Conference on Computer Vision*, 2016. 1, 2, 4, 6

[18] S. J. Oh, M. Fritz, and B. Schiele. Adversarial image perturbation for privacy protection – a game theory perspective. In *International Conference on Computer Vision*, 2017. 1, 2, 6

[19] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 2, 7

[20] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*, 1511.06434, 2015. 3, 4

[21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention(MICCAI)*, pages 234–241, 2015. 4

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[23] 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 *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016. 1, 2, 6

[24] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLRW*, 2014. 6

[25] Q. Sun, B. Schiele, and M. Fritz. A domain based approach to social relation recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 435–444, 2017. 4

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 6

[27] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3d shape from 2d motion. In *Advances in Neural Information Processing Systems*, pages 1555–1562, 2003. 3

[28] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. *arXiv*, 1705.00053, 2017. 2

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 5

[30] W. X and G. A. Generative image modeling using style and structure adversarial networks. *Journal of Foo*, 14(1):234–778, 2016. 2

[31] R. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv*, 1607.07539, 2016. 2

[32] A. Zadeh, T. Baltrusaitis, and L. Morency. Convolutional experts constrained local model for facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2051–2059, 2017. 3

[33] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. N. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[34] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. D. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813, 2015. 4