

# Shape Evasion: Preventing Body Shape Inference of Multi-Stage Approaches

Hosnieh Sattar<sup>1</sup> Katharina Krombholz<sup>2</sup> Gerard Pons-Moll<sup>1</sup> Mario Fritz<sup>2</sup>  
<sup>1</sup>Max Planck Institute for Informatics, <sup>2</sup>CISPA Helmholtz Center for Information Security  
 Saarland Informatics Campus, Saarbrücken, Germany,  
 {sattar,gpons}@mpi-inf.mpg.de, {fritz,krombholz}@cispa.saarland

## Abstract

Modern approaches to pose and body shape estimation have recently achieved strong performance even under challenging real-world conditions. Even from a single image of a clothed person, a realistic looking body shape can be inferred that captures a users' weight group and body shape type well. This opens up a whole spectrum of applications – in particular in fashion – where virtual try-on and recommendation systems can make use of these new and automatized cues. However, a realistic depiction of the undressed body is regarded highly private and therefore might not be consented by most people. Hence, we ask if the automatic extraction of such information can be effectively evaded. While adversarial perturbations have been shown to be effective for manipulating the output of machine learning models – in particular, end-to-end deep learning approaches – state of the art shape estimation methods are composed of multiple stages. We perform the first investigation of different strategies that can be used to effectively manipulate the automatic shape estimation while preserving the overall appearance of the original image.

## 1. Introduction

Since the early attempts to recognize human pose in images [64, 18], we have seen a transition to real-world applications where methods operate on challenging real-world conditions in uncontrolled pose and lighting. We have seen more recently progress towards extracting richer representations beyond the pose. Most notably, a full body shape that is represented by a 3D representation or a low dimensional manifold (SMPL) [32]. It has been shown that such representations can be obtained from fully clothed persons – even in challenging conditions from a single image [7] as well as from web images of a person [49].

On the one hand, this gives rise to various applications – most importantly in the fashion domain. The more accurate judgment of fit could minimize clothing returns, and avatars and virtual try-on may enable new shopping experiences.

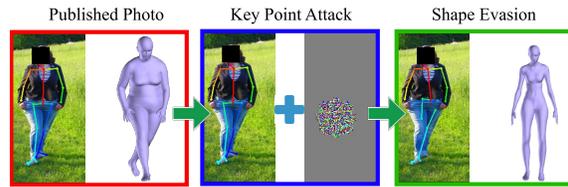


Figure 1: A realistic depiction of the undressed body is considered highly private and therefore might not be consented by most people. We prevent automatic extraction of such information by small manipulations of the input image that keep the overall aesthetic of the image.

Therefore, it is unsurprising that such technology already sees gradual adaption in businesses<sup>1</sup>, as well as start-ups<sup>2</sup>.

On the other hand, the automated extraction of such highly personal information from regular, readily available images might equally raise concerns about privacy. Images contain a rich source of implicit information that we are gradually learning to leverage with the advance of image processing techniques. Only recently, the first organized attempts were made to categorize private information in images [42] to raise awareness and to activate automatic protection mechanisms.

To control and control private information in images, a range of redaction and sanitization techniques have been introduced [41, 54, 58]. For example, evasion attacks have been used to disable classification routines to avoid extraction of information. Such techniques use adversarial perturbations to throw off a target classifier. It has been shown that such techniques can generalize to related classifiers [39], or can be designed under unknown/black-box models [36, 11, 8, 53, 27].

Unfortunately, such techniques are not directly applicable to state-of-the-art shape estimation techniques [7, 3, 31, 2, 21], as they are based on multi-stage processing. Typically, deep learning is used to extract person *keypoints*, and

<sup>1</sup><https://www.cnet.com/news/amazon-buys-body-labs-a-3d-body-scanning-tech-startup/>

<sup>2</sup><https://bodylabs.io/en/>

a model-fitting/optimization stage leads to the final keypoint estimation of pose and shape. As a consequence, there is no end-to-end architecture that would allow the computation of an image gradient needed for adversarial perturbations.

Today, we are missing successful evasion attacks on shape extraction methods. In this paper, we investigate to what extent shape extraction can be avoided by small manipulations of the input image. We follow the literature on adversarial perturbations and require our changes in the input image to be of a small Euclidean norm. After analyzing a range of synthetic attack strategies that operate at the keypoints level, we experimentally evaluate their effectiveness to throw off multi-stage methods that include a model fitting stage. These attacks turn out to be highly effective while leaving the images visually unchanged. In summary, our contributions are:

- An orientative user study of concerns w.r.t. privacy and body shape estimation in different application contexts.
- Analysis of synthetic attacks on 2D keypoint detections.
- A new localized attack on keypoint feature maps that require smaller noise norm for the same effectiveness.
- Evaluation of overall effectiveness of different attacks strategies on shape estimation. We show the first successful attacks that offer an increase in privacy with negligible loss in visual quality.

## 2. Related Works

This work relates to 3D human shape estimation methods, privacy, and adversarial image perturbation techniques. We will here cover recent papers in these three domains and some of the key techniques directly relating to our approach.

**Privacy and Computer Vision.** Recent developments in computer vision techniques [15, 30, 24, 37], increases concerns about extraction of private information from visual data such as age [6], social relationships [62], face detection [55, 60], landmark detection [68], occupation recognition [51], and license plates [70, 67, 10]. Hence several studies on keeping the private content in visual data began only recently such as adversarial perturbations [33, 43], automatic redaction of private information[41], predicting privacy risks in images [42], privacy-preserving video capture [1, 45, 35, 48], avoiding face detection [63, 22], full body re-identification [38] and privacy-sensitive life logging [25, 29]. None of the previous work in this domain studied the users shape privacy preferences. Hence, we present a new challenge in computer vision aimed at preventing automatic 3D shape extraction from images.

**3D Body Shape Estimation.** Recovery of 3D human shape from a 2D image is a very challenging task due to

ambiguities such as depth and unknown camera data. This task has been facilitated by the availability of 3D generative body models learned from thousands of scans of people [4, 47, 32], which capture anthropometric constraints of the population and therefore reduce ambiguities. Several works [49, 52, 20, 69, 12, 7, 26, 23, 69] leverage these generative models to estimate 3D shape from single or multiple images, using shading cues, silhouettes and appearance. Recent model based approaches are using deep learning based 2D detections [9] – by either fitting a model to them at test time [2, 49, 7, 3, 21] or by using them to supervise bottom-up 3D shape predictors [40, 44, 28, 59, 57, 2]. Hence, to evade recent shape estimators, we study different strategies to attack the 2D keypoint detections while preserving the overall appearance of the original image.

**Adversarial Image Perturbation.** Adversarial examples for deep neural networks were first reported in [56, 19] demonstrating that deep neural networks are being vulnerable to small adversarial perturbations. This phenomenon was analyzed in several studies [5, 65, 16, 50, 17], and different approaches have been proposed to improve the robustness of neural networks [43, 13]. Fast Gradient Sign Method (FGSM) and several variations of it were introduced in [19, 34] for generating adversarial examples that are indistinguishable—to the human eye—from the original image, but can fool the networks. However, these techniques do not apply to state of the art body shape estimation as those are based on multi-stage processing. Typically, shape inference consists in fitting a body model to detected skeleton keypoints. Consequently, we perturb the 2D keypoints to produce an error in the shape fitting step. Cisse et al. [14], proposed a method to fool 2D pose estimation. None of these techniques propose a solution to evade model based shape estimation. In order to evade 3D shape estimation in a subtle manner, we attack by removing and flipping individual keypoints. Since these attacks simulate typical failure modes of detectors (miss-detections due to occlusion and keypoint flips), they are more difficult to identify by the defender.

## 3. Understanding Users Shape Privacy Preferences

Modern body shape methods [49, 7, 40, 28] infer a realistic looking 3D body shape from a single photo of a person. The estimated 3D body captures user weight group and body shape type. However, such a realistic depiction of the undressed body is considered highly private and therefore might not be consented by most people. We performed a user study to explore the users’ personal privacy preferences related to their body shape data. Our goal was to study the degree to which various users are sensitive to sharing their shape data such as height, different body part measurement, and their 3D body shape in different contexts. This study

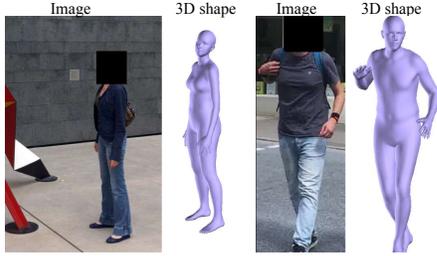


Figure 2: Participants were asked to indicate their comfort level for sharing these images publicly, considering they are the subject in these images.

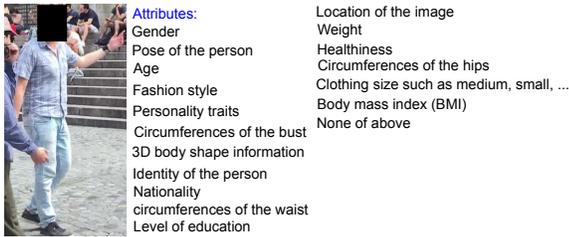


Figure 3: In Question 2 participants were shown this image, and were asked to select the attributes from the list that could be extracted.

was approved by our university’s ethical review board and is described next.

**User Study.** We split the survey into three parts. In the first part of the survey, our goal was to understand users image sharing preferences and the user’s knowledge of what type of information could be extracted from a single image.

*Part1-Question 1:* Users are shown Figure 2 without the 3D shape data. Participants are asked how comfortable they are sharing such images publicly, considering they are the subject in these images. Responses are collected on a scale of 1 to 5, where: (1) Extremely comfortable, (2) Slightly comfortable, (3) Somewhat comfortable, (4) Not comfortable, and (5) Extremely uncomfortable.

*Part1-Question 2:* Participants were shown Figure 3, and were asked which attributes could be extracted from this image.

In the second part, users were introduced to 3D shape models by showing them images of 8 people along with their 3D body shape, as shown in Figure 4. The purpose of part 2 was to understand the user’s perceived closeness of extracted 3D shapes to the original images, and their level of comfort with them.

*Part2-Question 3:* Participants were asked to rate how close the estimated 3D shape is to the person in the image. Responses are collected on a scale of 1 to 5, where: (1) Untrue of the person in the image, (2) somewhat untrue of the person in the image, (3) Neutral, (4) Somewhat true of



Figure 4: Participants were asked to judge the closeness of the depicted 3D shape to the actual body of the person in the images.

the person in the image, and (5) True of the person in the image.

*Part2-Question 4:* Participants were shown Figure 2 asked to indicate how comfortable they are sharing such a photograph along with 3D shape data publicly, considering they are the subject in these images. We collected responses on a scale of 1 to 5, similar to Question 1.

In the third part of this survey, we explore users preferences on what type of body shape information they would share for applications such as (a) Health insurance, (b) Body scanners at airport, (c) Online shopping platforms, (d) Dating platforms, and (e) Shape tracking applications ( for sport, fitness, ...).

*Part3-Question 5:* Users were asked their level of comfort on a scale of 1 to 5 for the applications mentioned above.

**Participants.** We collect responses of 90 unique users in this survey. Participants were not paid to take part in this survey. Out of the 90 respondents, 43.3% were female, 55.6% were male, and 1.1% were queer. The dominant age range of our participants (63.3%) was in 21-39, followed by 30-39 (23.3%). Participants have a wide range of education level, where 46.7% has master degree, 21.1% has bachelor degree<sup>3</sup>.

**Analysis.** The results of *Part1-Question 1* and *Part2-Question 4* are shown in Figure 5a. We see that majority of the users do not feel comfortable or they feel extremely uncomfortable (36%, 30%) sharing their 3D data publicly compared to sharing only their images (29%, 14%).

In *Part1-Question 2*, the top three selected attributes were: gender (98.9%), pose (87.8%), and age (85.6%). Shape related attributes such as body mass index (BMI) (47.8%), weight (63.3%), and 3D body shape (66.7%) were

<sup>3</sup>Further details on participants demographic data are presented in the supplementary materials.

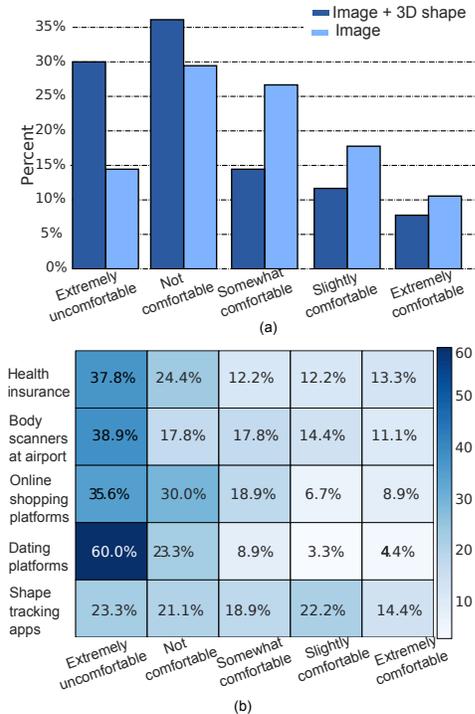


Figure 5: (a) Comfort level of participants in sharing images with and without 3D mesh data, considering they are the subject in these images. (b) Comfort level of the participant for sharing their 3D mesh data with multiple applications. Results are shown as the percentage of times an answer is chosen.

not in the top selected attributes, indicating that many participants were unaware that such information could be extracted from an image using automatic techniques.

In *Part3-Question3*, users were asked to judge the quality of the presented 3D models. Around 43% of the participants believe the presented shape is Somewhat true of the person in the image, and 31% thinks the 3D mesh is true to the person in the picture. This indicates that recent approaches can infer perceptually faithful 3D body shapes under clothing from a single image.

Figure 5b presents the results from *Part3-Question 5*. Participants show a high level of discomfort in sharing their 3D shape data for multiple applications. In all investigated applications except for fitness, the majority of the users responded with "discomfort of some degree".

The user study demonstrates that users are concerned about the privacy of their body shape. Consequently, we present next a framework to prevent 3D shape extraction from images.

## 4. Shape Evasion Framework

Model-based shape estimation methods from a 2D image are based on a two-stage approach. First, a neural network

is used to detect a set of 2D body keypoints, then a 3D body model fits the detected keypoints. Since this approach is not end-to-end, it does not allow direct computation of the image gradient needed for adversarial perturbation. To this end, we approach the shape evasion by attacking the keypoints detection network. In section 4.1, we give a brief introduction on model-based shape estimation method. In section 4.2, we introduced a local attack that allows targeted attacks on keypoints. Figure 6 shows an overview of our approach.

### 4.1. Model Based Shape Estimation

A Skinned Multi-Person Linear Model (SMPL) [32] is a state of the art generative body model. The SMPL function  $M(\beta, \theta)$ , uses shape  $\beta$  and poses  $\theta$  to parametrize the surface of the human body that is represented using  $N = 6890$  vertices. The shape parameters  $\beta \in \mathbb{R}^{10}$  encode changes in height, weight and body proportions. The body pose  $\theta \in \mathbb{R}^{3P}$ , is defined by a skeleton rig with  $P = 24$  keypoints. The 3D skeleton keypoints are predicted from body shape via  $J(\beta)$ . We can use a global rigid transform  $R_\theta$  to pose the SMPL keypoint. Hence,  $R_\theta(J(\beta)_i)$  denotes a posed 3D keypoint  $i$ . In order to estimate 3D body shape from a 2D image  $I$ , several works [49, 7, 31], minimize an objective function composed of a keypoint-based data term, pose priors, and a shape prior.

$$E(\beta, \theta) = E_{P_\theta}(\theta) + E_{P_\beta}(\beta) + E_J(\beta, \theta; \mathbf{K}, \mathbf{J}_{\text{est}}) \quad (1)$$

where  $E_{P_\theta}(\theta)$ , and  $E_{P_\beta}(\beta)$  are the pose and shape prior terms as described in [7]. The  $E_J(\beta, \theta; \mathbf{K}, \mathbf{J}_{\text{est}})$  is the keypoint-based data term which penalizes the weighted 2D distance between estimated 2D keypoints,  $\mathbf{J}_{\text{est}}$ , and the projected SMPL body keypoint  $R_\theta(J(\beta))$ :

$$E_J(\beta, \theta; \mathbf{K}, \mathbf{J}_{\text{est}}) = \sum_{\text{keypoint } i} w_i \rho(\Pi_{\mathbf{K}}(R_\theta(J(\beta)_i)) - \mathbf{J}_{\text{est}, i}) \quad (2)$$

where  $\Pi_{\mathbf{K}}$  is the projection from 3D to 2D of the camera with parameters  $\mathbf{K}$  and  $\rho$  a Geman-McClure penalty function which is robust to noise in the 2D keypoints detections.  $w_i$  indicates the confidence of each keypoints estimate, provided by 2D detection method. For cases such as occluded or missing keypoints,  $w$  is very low, and hence the data term will be driven by pose prior term. Furthermore, the prior term avoids impossible poses. Shape evasion can be achieved by introducing error in 2D keypoints detection  $\mathbf{J}_{\text{est}}$ . We use Adversarial perturbation to fool the pose detection method by either removing a keypoint or flipping two keypoints with each other.

### 4.2. Adversarial Image Generation

The state of the art 2D pose detection methods use a neural network  $f$  parametrized by  $\phi$ , to predict a set of

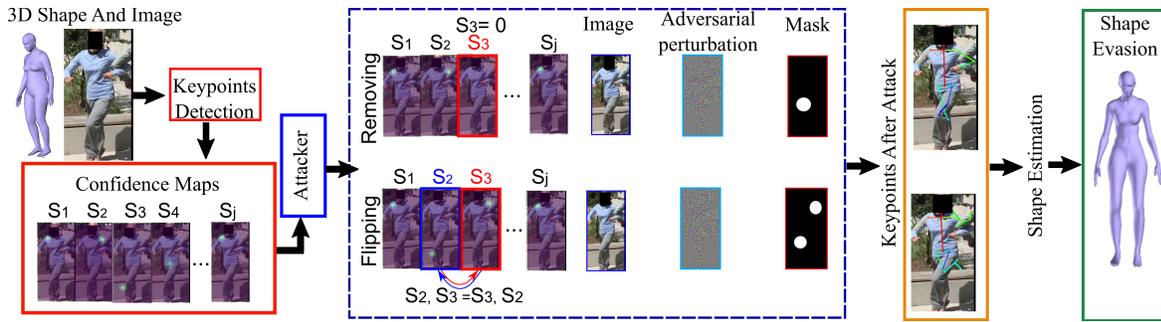


Figure 6: The summary of our framework. We assume that we have full access to the parameter of the network. The attacker breaks the detections by removing or flipping of a keypoint. Hence the final estimated shape does not depict the person in the image.

2D locations of anatomical keypoints  $\mathbf{J}_{est}$  for each person in the image. The network produces a set of 2D confidence maps  $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \dots, \mathbf{S}_P\}$ , where  $\mathbf{S}_i \in \mathbb{R}^{w \times h}$ ,  $i \in 1, 2, 3, \dots, P$ , is a confidence map for the keypoints  $i$  and  $P$  is total number of Keypoints. Assuming that a single person is in the image, then each confidence map contains a single peak if the corresponding part is visible. The final set of 2D keypoints  $\mathbf{J}_{est}$  are achieved by performing non-maximum suppression per each confidence map. These confidence maps are shown in Figure 6.

To attack a keypoint we used adversarial perturbation. Adding adversarial perturbation  $\mathbf{a}$  to an image  $I$  will causes a neural network to change its prediction[56].The adversarial perturbation  $\mathbf{a}$  is defined as the solution to the optimization problem

$$\arg \min_{\mathbf{a}} \|\mathbf{a}\|_2 + L(f(I + \mathbf{a}; \phi), \mathbf{S}^*). \quad (3)$$

$L$  is the loss function between the network output and desired confidence maps  $\mathbf{S}^*$ .

**Removing and Flipping of Keypoints:** The  $\mathbf{S}^*$  is defined for removing and flipping of keypoints. to remove a keypoint, we put its confidence map to zero. For example if we are attacking the first keypoint we have:  $\mathbf{S}^* = \{\mathbf{S}_1 = 0, \mathbf{S}_2, \mathbf{S}_3, \dots, \mathbf{S}_P\}$ . To flip two key points we exchanged the values of two confidence map as  $\mathbf{S}^i, \mathbf{S}^j = \mathbf{S}^j, \mathbf{S}^i$ . In case  $i, j = 2, 3$  we have  $\mathbf{S}^* = \{\mathbf{S}_1, \mathbf{S}_3, \mathbf{S}_2, \dots, \mathbf{S}_P\}$ . An example of removing and flipping of the keypoint is shown in Figure 6.

**Fast Gradient Sign Method (FGSM) [19]:** FGSM is a first order optimization schemes used in practice for Equation 3, which approximately minimizes the  $\ell_\infty$  norm of perturbations bounded by the parameter  $\epsilon$ . The adversarial examples are produced by increasing the loss of the network on the input  $I$  as

$$I^{adv} = I + \epsilon \text{sign}(\nabla_I L(f(I; \phi), \mathbf{S}^*)). \quad (4)$$

We call this type of attack global as the perturbation is applied to the whole image. This perturbation results in poses with several missing keypoints or poses outside of natural human pose manifold. While this will often make the subsequent shape optimization step fail (Eq. (2)), the approach has two limitations: i) this attack requires a large perturbation and ii) the attack is very easy to identify by the defender.

**Masked Fast Gradient Sign Method (MFGSM):** To overcome the limitations of the global approach, we introduced Masked FGSM. This allows for localized perturbation for more targeted attacks. This method will generate poses, which are close to ground truth pose, yet have a missing keypoint that will cause shape evasion—while requiring smaller perturbations as shown in the experiments. We will refer to this scheme as “local” in the rest of the paper. To attack a specific keypoint we solve the following optimization problem in a iterative manner as:

$$I_0^{adv} = I$$

$$I_{t+1}^{adv} = \text{clip}(I_t^{adv} - \alpha \cdot \text{sign}(\nabla_{I_t^{adv}} L(f(I_t^{adv}; \phi), \mathbf{S}^*) \odot M), \epsilon) \quad (5)$$

where mask  $\mathbf{M} \in \mathbb{R}^{w \times h}$  is used to attack a keypoint  $\mathbf{J}_{est,i} \in \mathbb{R}^2$  selectively.  $\mathbf{M}$  is defined as:

$$\mathbf{M} = \begin{cases} 1 & \text{if } (\mathbf{x} - \mathbf{J}_{est,i})^2 = r^2 \\ 0 & \end{cases}$$

$r$  controls the spread of the attack and  $\mathbf{x} \in \mathbb{R}^2$  are the pixel coordinates. To ensure the max norm constraint of perturbation  $\mathbf{a}$  being no greater than  $\epsilon$  is preserved, the  $\text{clip}(z, \epsilon)$  is used, which keeps the values of  $z$  in the range  $[z - \epsilon, z + \epsilon]$ .

## 5. Experiments

The overall goal of the experimental section is to provide an understanding and the first practical approach to evade body shape estimation and hence protect the privacy of the

Attack	Right ankle	Right knee	Right hip	Left hip	Left knee	Left ankle	Right wrist	Right elbow	Right shoulder	Left shoulder	Left elbow	Left wrist	Head top	Average
Real	<b>1.32</b>	<b>1.4</b>	1.39	1.37	<b>1.38</b>	<b>1.32</b>	<b>1.36</b>	<b>1.41</b>	1.40	1.35	<b>1.28</b>	<b>1.37</b>	<b>1.35</b>	<b>1.37</b>
Synthetic	1.17	1.18	<b>1.79</b>	<b>1.94</b>	1.18	1.18	1.18	1.17	<b>1.43</b>	<b>1.49</b>	1.15	1.16	1.19	1.32

Table 1: Shape estimation error on 3DPW with Procrustes analysis with respect to the ground truth shape. Error in cm. The goal of each attack is to induce bigger error in the estimated shape. Hence, higher errors are indication of a successful attack.

user. We approach this by systematically studying the effect of attacking keypoint detections on the overall shape estimation pipeline. First, we study synthetic attacks based on direct manipulation of keypoints locations, where we can observe the effects on body shape estimation in an idealized scenario. This study is complemented by real image-based attacks which make keypoint estimation fail. Together, we evaluate our approach that provides the first and effective defence against body shape estimation on real-world data.

**Dataset.** We used 3D Poses in the Wild Dataset (3DPW) [61], which includes 60 sequences with 7 actors. To achieve ground truth shape parameter  $\beta$ , actors were scanned and SMPL was non-rigidly fit to them to obtain their 3D models similar to [46, 66]. To the best of our knowledge, 3DPW is the only in wild image dataset which provides the ground truth shape data as well, which makes this dataset most suitable for our evaluation. For our evaluation, for each actor, we randomly selected multiple frames from different sequences. All reported results are averaged across subjects and sampled sequence frames– the exact sampled frames are specified in the supplementary material.

**Model.** We used OpenPose [9] for keypoint detection as it is the most widely used. OpenPose consists of a two-branch multi-stage CNN, which process images at multi-scales. Each stage in the first branch predicts the confidence map  $S$ , and each stage in the second branch predicts the Part Affinity Fields (PAFs). For the shape estimation, we used the public code of Smplify [7], which infers a 3D mesh by fitting the SMPL body model to a set of 2D keypoints. To improve the 3D accuracy, we refined the estimations using silhouette as described in [49]. We used MFGSM (Eq. (5) with  $\alpha = 1$ ) in an iterative manner. We evaluated attacks when setting the  $\ell_\infty$  norm of the perturbations to  $\epsilon = 0.035$  since we observed that higher values lead to noticeable artifacts in the image. We stop the iterations if we reach an Euclidean distance (between the original and perturbed images) of 0.02 in image space for local, and 0.04 for global attacks.

### 5.1. Synthetic Modification of The keypoints

First, we studied the importance of each keypoint on the overall body shape estimation by removing one keypoint at a time–which simulate miss-detections. The error on

shape estimation caused by this attack is reported in the second row of Table 1. We observe that removing “Hips”, and “Shoulder” keypoints results in the highest increase of error of 34%, and 25.86% whereas “Elbows” and “Wrists” result in an increase of only 1%.

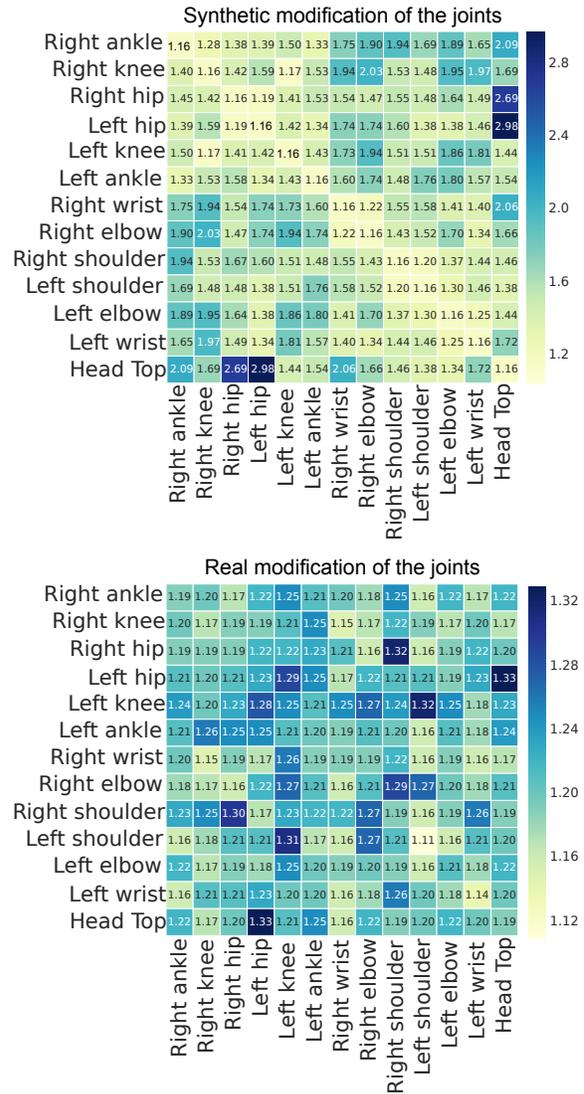


Figure 7: Shape estimation error on 3DPW with Procrustes analysis. Error in cm for synthetic and real flipping of the keypoints.

We also studied the effect of flipping keypoints. The results of this experiment are shown in Figure 7. Flipping the “Head” with the left or right “Hip” caused an increase in er-

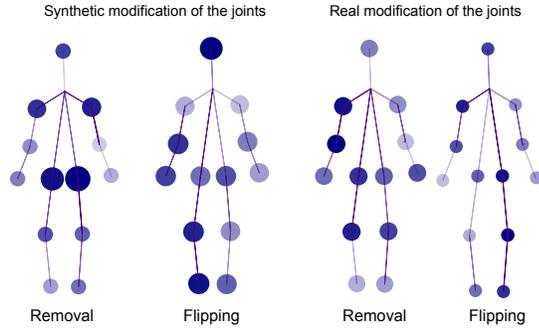


Figure 8: The overall shape estimation error induced by synthetic and real (local) attacks. The darker and bigger circles shows higher error.

ror of 143.96%. Flipping the “Elbow” and “Knee” was the second most effective attack causing 67% increase of error in average. The least effective attack was by flipping the left and right knee (2.58%). The average error introduced by removing or flipping of each keypoint is illustrated in Figure 8 – higher error is larger in size and darker in colour. We can see that, overall “Hip”, “Shoulder”, and “Head” keypoints play a crucial role in the quality of the final estimated 3D mesh, and are the most powerful attacks.

## 5.2. Attacking keypoint Detection by Adversarial Image Perturbation

To apply modifications to the keypoints, we used our proposed local Mask Iterative Fast Gradient Sign Method (MIFGSM). Figure 9 shows the keypoint confidence map values when removing and adding a keypoint using local and global attacks with respect to the amount of perturbation added to the image. We can see that the activation’s per each keypoint decreases after each iteration. Interestingly, the rate of decrease is slower for global attacks for the same amount of perturbation (0.015 Mean Squared Error (MSE) between perturbed and original image). Global attacks require much higher amount of perturbations (0.035 MSE) to be successful, causing visible artifact in the image. We observed similar behavior when adding a “fake” keypoint detections (required to flip two keypoints). Similarly, the rate of increase in activation was slower for global compared to local for the same amount of perturbation (0.015 MSE, blue bar in the plot). From Figure 9 we can also see that shoulders and head are more resistant to the removal. Furthermore, the attack was the most successful in the creation of wrists. Since local attacks are more effective, we consider only the local attack method for further analysis.

## 5.3. Shape Evasion

In this section, we evaluate the effectiveness of the whole approach for evading shape estimation and therefore, protecting the users’ privacy. We used our proposed local

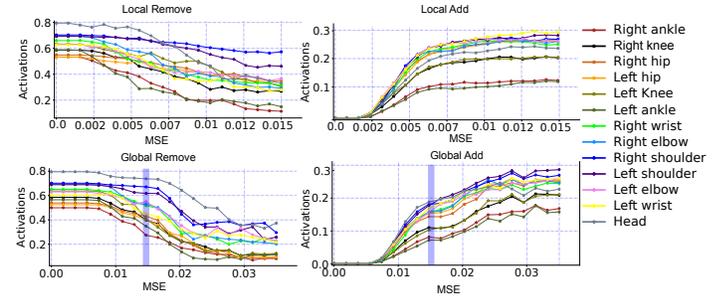


Figure 9: Comparison of local and global attacks for removing and adding a keypoint. The local attack has a higher rate of decrease or increase of activation compared to the global method for the same amount of perturbation. The blue bar on the global plots shows the end of the local methods.

method to remove and flip keypoints instead of the synthetic modification of the keypoints as described in section 5.1. Hence, we call this attack as a real modification of keypoints.

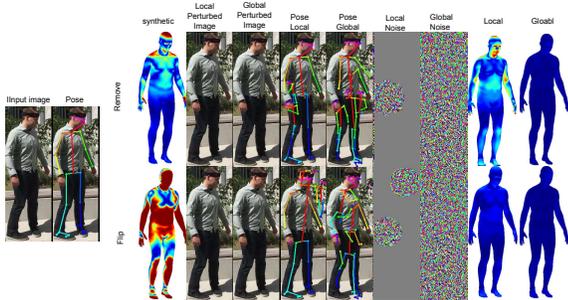
The error on shape estimation caused by removing of the keypoints using our local method are reported in the first row of Table 1, we refer to it as real. We see that attacks on “Right Elbow” and “Right knee” causes 21% increase of error in shape estimation. The least amount of error 10% and 13% was produced by removing “Left Elbow” and “Ankles” respectively. However, “Hip” and “Shoulder” gained higher error in average for left and right keypoint by 18%. On average, the real attack for removing keypoints caused an even higher error than the synthetic mode (18% to 13%), showing the effectiveness of this approach in shape evasion and hence protecting the users’ privacy.

The result for flipping the keypoints is shown in Figure 7 (Real modification of the keypoint). The highest increase in error was (14%) caused by flipping the “Head” with “Left Hip”, the second most effective attack was for flipping the “Shoulder” and “Knee” keypoints (12% in average over left and right keypoints). The least effective attack was on “wrist” with increase of 2% error on average.

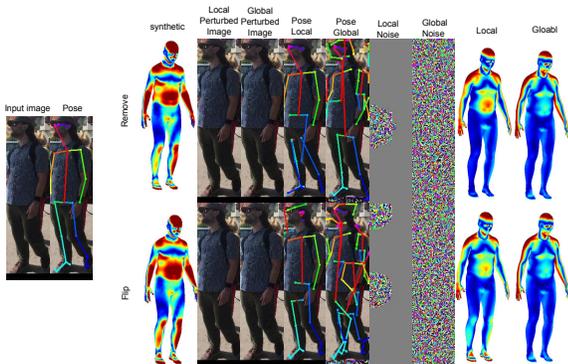
Real flipping of keypoints achieves an error of (3%) compared to real removing attacks (18%), which shows they are slightly less effective. In addition, similar to global attacks, flipping of keypoints causes more changes in the keypoints, making the detection of these attacks easier.

## 5.4. Qualitative Results

In Figure 10, we present example results obtained for each type of attack. The global attack causes pose estimation to hallucinate multiple people in the image, destroying the body signal of the person in the picture. As the predicted poses in the global attack are not in human body manifold, the optimization step in SMPL will fail to fit these keypoints resulting in average shape estimates. In the local attack, we were able to apply small changes in the keypoints. Hence,



(a) Person with body shape close to SMPL template (0.04 cm)



(b) Person with a higher distance to SMPL template (2 cm)

Figure 10: The left side shows the original image with the estimated pose, and the right the output when modified with local and global adversarial perturbations with corresponding error heatmaps with respect to ground truth shapes (red means  $> 2\text{cm}$ ). Here we applied local and global attack for removing the “Right Hip”, and flipping the “Right Hip” and ‘Head Top’. The global attack causes the pose estimation to hallucinate multiple people in the image, while or local attack only changes the selected keypoints. The predicted shape in case of a global attack is always close to the average template of SMPL causing a lower error for people with an average shape.

these small changes make the shape optimization stage predict shapes that are not average and also not close to the person in the image. Overall, shape evasion was most successful when removing the keypoints than flipping them, and when using the local attacks.

## 6. Discussion

As our study of privacy on automatically extracted body shapes and method for evading shape estimation is the first of its kind, it serves as a starting point – but naturally needs further investigations to extend on both lines of research that we have touched on. The following presents a selection of open research questions.

**Targeted vs untargeted shape evasion.** While our method for influencing the keypoints detection is targeted, the overall approach to shape evasion remains untargeted. Depending on the application scenario, a consistent change or particular randomization of the change in shape might be desired, which is not addressed by our work.

**Effects of adversarial training.** It is well known that adversarial training against particular image perturbations can lead to some robustness against such attacks [56, 39] and in turn, the attack can again be made to some extent robust against such defences. Preventing this cat-mouse-game is subject of on-going research and – while very important – we consider outside of the scope of our first demonstration of shape evasion methods.

**Scope of the user study.** While our user study encompasses essential aspects of privacy of body shape information, clearly a more detailed understanding can be helpful to inform the design evasion techniques and privacy-preserving methodologies that comply with the users’ expectations on handling personal data. As our study shows that such privacy preferences are personal as well as application domain specific, there seem ample opportunities to leverage the emerging methods of high-quality body shape estimation in compliance with user privacy.

## 7. Conclusion

Methods for body shape estimation from images of clothed people are getting more and more accurate. Hence we have asked the timely question to what extent this raises privacy concerns and if there are ways to evade shape estimation from images. To better understand the privacy concerns, we conduct a user study that sheds light on the privacy implication as well as the sensitivity of shape data in different application scenarios. Overall, we observe a high sensitivity, which is also dependent on the use case of the data. Based on this understanding, we follow up with a defence mechanism that can hamper or even prevent body shape estimation from real-world images. Today’s state of the art body shape estimation approaches are frequently optimization based and therefore don’t lend themselves to gradient-based adversarial perturbation. We tackle this problem by a two-stage approach that first analysis the effect of individual keypoints on the shape estimate and then proposes adversarial image perturbations to influence the keypoints. In particular, our novel localized perturbation techniques constitute an effective technique to evade body shape estimation at negligible changes to the original image.

## References

- [1] P. Aditya, R. Sen, P. Druschel, S. Joon Oh, R. Benenson, M. Fritz, B. Schiele, B. Bhattacharjee, and T. T. Wu. I-pic: A platform for privacy-compliant image capture. *ACM*, 2016.
- [2] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*, 2019.
- [3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *CVPR Spotlight*, 2018.
- [4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: shape completion and animation of people. In *ACM Transactions on Graphics*, 2005.
- [5] A. Arnab, O. Miksik, and P. H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, 2018.
- [6] C. Bauckhage, A. Jahanbekam, and C. Thurau. Age recognition in the wild. In *ICPR*, 2010.
- [7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016.
- [8] W. Brendel and M. Bethge. Comment on "biologically inspired protection of deep networks from adversarial attacks". *arxiv*, 1704.01547, 2017.
- [9] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.
- [10] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen. Automatic license plate recognition. *IEEE Trans. Intelligent Transportation Systems*, 5:42–53, 2004.
- [11] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. *ACM*, 2017.
- [12] Y. Chen, T.-K. Kim, and R. Cipolla. Inferring 3d shapes and deformations from single views. In *ECCV*, 2010.
- [13] M. Cissé, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *ICML*, 2017.
- [14] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *NIPS*, 2017.
- [15] J. Deng, W. Dong, R. Socher, L. jia Li, K. Li, and L. Feifei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [16] A. Fawzi, O. Fawzi, and P. Frossard. Analysis of classifiers? robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.
- [17] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard. Robustness of classifiers: from adversarial to random noise. In *NIPS*, 2016.
- [18] D. Gavrilu. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73:82–98, 1999.
- [19] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arxiv*, abs/1412.6572, 2014.
- [20] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009.
- [21] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics*, (Proc. SIGGRAPH), jul 2019.
- [22] A. Harvey. Cv dazzle: Camouflage from computer vision. *Technical report*, 2012.
- [23] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H.-P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *CVPR*, 2010.
- [24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [25] R. Hoyle, R. Templeman, D. Anthony, D. Crandall, and A. Kapadia. Sensitive lifelogs: A privacy analysis of photos from wearable cameras. In *CHI*, 2015.
- [26] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *3DV*, 2017.
- [27] A. Ilyas, L. Engstrom, and A. Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018.
- [28] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.
- [29] M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia. Enhancing lifelogging privacy by detecting screens. In *CHI*, 2016.
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [31] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017.
- [32] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, oct 2015.
- [33] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In *CVPR*, 2007.
- [34] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deep-fool: A simple and accurate method to fool deep neural networks. In *CVPR*, 2016.
- [35] C. Neustaedter, S. Greenberg, and M. Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. 2006.
- [36] S. J. Oh, M. Augustin, B. Schiele, and M. Fritz. Towards reverse-engineering black-box neural networks. In *ICLR*, 2018.
- [37] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Person recognition in personal photo collections. In *ICCV*, 2015.
- [38] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele. Faceless person recognition: Privacy implications in social media. In *ECCV*, 2016.

- [39] S. J. Oh, M. Fritz, and B. Schiele. Adversarial image perturbation for privacy protection – a game theory perspective. In *ICCV*, 2017.
- [40] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018.
- [41] T. Orekondy, M. Fritz, and B. Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *CVPR*, 2018.
- [42] T. Orekondy, B. Schiele, and M. Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *ICCV*, 2017.
- [43] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 2016.
- [44] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018.
- [45] F. Pittaluga and S. J. Koppal. Privacy preserving optics for miniature vision sensors. In *CVPR*, 2015.
- [46] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics*, 36(4), 2017.
- [47] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34:120, 2015.
- [48] N. Raval, A. Srivastava, K. Lebeck, L. Cox, and A. Machanavajjhala. Markit: Privacy markers for protecting visual secrets. In *UbiComp*, 2014.
- [49] H. Sattar, G. Pons-Moll, and M. Fritz. Fashion is taking shape: Understanding clothing preference based on body shape from online sources. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 968–977. IEEE, 2019.
- [50] U. Shaham, Y. Yamada, and S. Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arxiv*, abs/1511.05432, 2015.
- [51] M. Shao, L. Li, and Y. Fu. What do you do? occupation recognition in a photo via social context. In *CVPR*, 2013.
- [52] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*. 2008.
- [53] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *arxiv*, abs/1710.08864, 2017.
- [54] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele. A hybrid model for identity obfuscation by face replacement. In *ECCV*, 2018.
- [55] X. Sun, P. Wu, and S. C. H. Hoi. Face detection using deep learning: An improved faster rcnn approach. *CoRR*, abs/1701.08289, 2017.
- [56] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.
- [57] V. Tan, I. Budvytis, and R. Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, 2017.
- [58] E. Tretschk, S. J. Oh, and M. Fritz. Sequential attacks on agents for long-term adversarial goals. In *2. ACM Computer Science in Cars Symposium – Future Challenges in Artificial Intelligence and Security for Autonomous Vehicles*, 2018.
- [59] H. Tung, H. Wei, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017.
- [60] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2001.
- [61] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, sep 2018.
- [62] G. Wang, A. C. Gallagher, J. Luo, and D. A. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, 2010.
- [63] M. J. Wilber, V. Shmatikov, and S. Belongie. Can we still avoid automatic face detection? In *WACV*, 2016.
- [64] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfunder: real-time tracking of the human body. *TPAMI*, 19(7):780–785, July 1997.
- [65] X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darell, and D. Song. Can you fool ai with adversarial examples on a visual turing test? *CoRR*, abs/1709.08693, 2017.
- [66] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *CVPR*, 2017.
- [67] H. Zhang, W. Jia, X. He, and Q. Wu. Learning-based license plate detection using global and local features. In *ICPR*, 2006.
- [68] Y. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *CVPR*, 2009.
- [69] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. In *ACM Transactions on Graphics*, 2010.
- [70] W. Zhou, H. Li, Y. Lu, and Q. Tian. Principal visual word discovery for automatic license plate detection. *IEEE Trans. Image Processing*, 21:4269–4279, 2012.