

On the Robustness of Human Pose Estimation

¹Naman Jain[†] ¹Sahil Shah[†] ²Abhishek Kumar ¹Arjun Jain

¹Department of Computer Science, IIT Bombay, ²Gobasco AI Labs
{namanjain, sahilshah, ajain}@cse.iitb.ac.in, abhisharaya@gmail.com

Abstract

This paper provides, to the best of our knowledge, the first comprehensive and exhaustive study of adversarial attacks on human pose estimation. Besides highlighting the important differences between well-studied classification and human pose-estimation systems w.r.t. adversarial attacks, we also provide deep insights into the design choices of pose-estimation systems to shape future work. We compare the robustness of several pose-estimation architectures trained on the standard datasets, MPII and COCO. In doing so, we also explore the problem of attacking non-classification based networks including regression based networks, which has been virtually unexplored in the past.

We find that compared to classification and semantic segmentation, human pose estimation architectures are relatively robust to adversarial attacks with the single-step attacks being surprisingly ineffective. Our study show that the heatmap-based pose-estimation models fare better than their direct regression-based counterparts and that the systems which explicitly model anthropomorphic semantics of human body are significantly more robust. We find that the targeted attacks are more difficult to obtain than untargeted ones and some body-joints are easier to fool than the others. We present visualizations of universal perturbations to facilitate unprecedented insights into their workings on pose-estimation. Additionally, we show them to generalize well across different networks on both the datasets.

1. Introduction

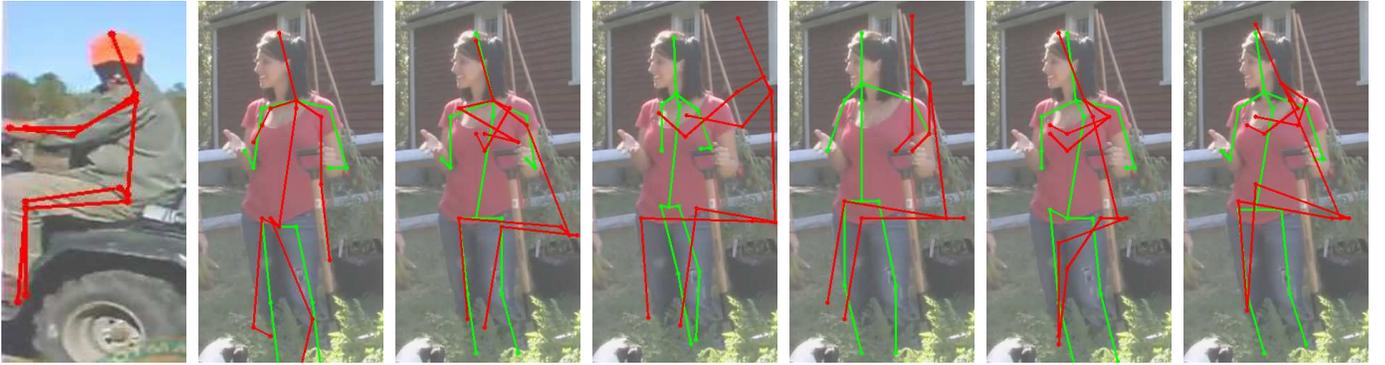
The past few years have witnessed an exponential growth in the real-world deployment of deep-learning based automation systems, due to its phenomenal ability to learn complex task-dependent features and decision-functions directly from the data. However, alongside their innumerable successes deep-learning systems are extremely prone to adversarial attacks which refer to imperceptible noise that can significantly affect performance! Therefore, the study and

defense against adversarial attacks on deep-learning systems is critical towards their real-world deployment.

Discovering the extent of the harmful effects of adversarial examples is still and active area of research. The study of adversarial attacks on classification systems [2, 4, 6, 11, 12, 20, 22, 29, 30] has seen more activity than regression systems [7, 34]. Human-pose estimation, referred to as **HPE** for brevity, is one such application that uses a blend of regression and classification approaches to learn the compositionality of human bodies, warranting a separate study. To this end, we present the first comprehensive study of the effects of adversarial attacks on HPE systems and their effectiveness with respect to different design choices like heatmaps vs. direct regression, multi-scale processing, attention and compositional constraints.

Our analysis on two standard datasets, MPII [1] and COCO [21], yields interesting insights that could prove useful for shaping the future of robust deep-learning based HPE systems. Our studies show that heatmap-based approaches are more robust than direct joint-regression and among the former, the networks that model compositional human constraints are more robust. We also find that imagenet pre-training improves the robustness of network. We observe that HPE networks are more difficult to attack than their classification counterparts. Among targeted and un-targeted attacks, the former are harder to obtain and also require carefully tuned hyper-parameters. We also provide a thorough study of adversarial attacks on the most popular HPE backbone, Stacked Hourglass [26], and show that an attack on features deep within the model is far more detrimental than just on the final output. Then we show that universal adversarial perturbations [15, 24] are detrimental to HPE systems and supplement this finding with their visualizations which hallucinate body-joints. We show that the universal perturbations generalize fairly well across networks that makes them a serious threat to HPE systems. Our analysis on the vulnerability of different joints towards adversarial attacks reveal that the hip and the joints below the hip are the most vulnerable while head and neck are most stable. Lastly we also test some image-processing techniques on adversarial examples and show their effects

[†]equal contribution



(a) Target Pose (b) Attention-HG [8] (c) 8-Stacked-HG [26] (d) DeepPose [39] (e) Chained-Preds [13] (f) DLCM [37]Ta (g) 2-Stacked-HG [26]
 Figure 1. Example of various targeted adversarial attacks of different networks on the MPII benchmark. (a) represents the target pose used for computing the adversarial perturbation while in figures (b-g) : Green skeletons show the original predictions while the red skeletons show the predictions for perturbed image. For more visualizations refer supp. mat.

2. Related Work

Soon after AlexNet that made deep neural networks, DNNs for brevity, popular, [36] showed that DNNs are easily fooled by noise computed using L-BFGS technique. Later, [14] introduced Fast-Gradient-Sign-Method (FGSM) that was more efficient using only gradient ascent instead of L-BFGS. Then, [25] introduced Iterative-Gradient-Sign-Method (IGSM) and [19] made it stronger by optimizing for the least likely class. Since then there has been a lot of work in this field that extended these attacks with different datasets, penalty functions and optimization methods [4, 5, 6, 9, 11, 20, 22, 25, 27, 29, 35]. An altogether different line of work employed DNNs to directly generate adversarial perturbations from an input image [3, 30, 33, 41]. These approaches require complete access to the network limiting their practicality for real-world application. Black-box attacks [22, 28, 29] generalize across networks and do not need access to the target network that makes them more practical.

Most of the aforementioned attacks are image-specific and need costly back-propagation through the entire network. To mitigate this issue, a universal adversarial perturbation [15, 24] can be obtained for a DNN that can be added to any image to fool the network. [24] show the effectiveness of universal attacks on the ImageNet, while [15] analyzed the same for semantic segmentation. Mostly, the study of adversarial attacks has been limited to image classification, only recently, they have been analyzed in other settings such as image segmentation (again a per-pixel *classification*) [2, 12, 15, 30, 40, 42], object detection [7, 34], visual question answering [43].

For human pose, on the other hand, there hasn't been much study of adversarial attacks and the closest work to ours is [9] that explores metric specific loss functions for different tasks. Their focus was on exploiting loss function frameworks to develop metric specific attacks and they

demonstrate their approach on classification, segmentation and HPE. Therefore, their study on HPE does not cover it in detail rather showcases it as application of their generic framework. We, on the other hand, present a comprehensive analysis of adversarial attack on the HPE systems to obtain deeper insights.

3. Background, Notations and Experimental Settings

This section contains background on HPE and adversarial attack to facilitate understanding and the details of experimental settings with notations.

3.1. Human Pose Estimation (HPE)

It refers to inferring a set of 2D joint-locations or pose, $P = \{P_1, P_2, \dots P_k\}$ for k body joints from an input RGB image, I , that contains a human. The first DNN based approach, DeepPose [39], used AlexNet [18] followed by *direct regression* for ground-truth \hat{P} from I . Later, [38] introduced heatmaps that represents k joint-locations with the help of k channels, one for each joint, with Gaussian bumps centered at the corresponding joint locations. The input image, I , is passed through multiple resolution banks and multi-scales features from different resolutions are concatenated to regress for the heatmaps. In [26], the authors introduced a recurring structure that feeds the previously predicted heatmaps for further processing with image features, referred to as Stacked Hourglasses, it has been used as the backbone architecture in numerous works and led to significant improvement in the performance over previous approaches. In order to provide a comprehensive coverage of HPE systems for our study we analyze five different architectures.

DeepPose and **Stacked-Hourglass** or **SHG** for brevity [26], are already explained in the paragraph above. We used two different variants with 2 and 8 stacks termed as **2-SHG**

and **8-SHG. Chained-Prediction** [13] casts HPE as a sequential joint prediction with a series of encoder-decoder networks that predict heatmaps of joints, thus conditioning the prediction of joints over the pre-computed joints. **Hourglass Attention** [8] incorporates multi-context attention by utilizing CRFs to model the correlations between neighbouring regions in the attention map. **Deeply-Learned-Compositional-Model or DLCM** [37] uses hourglass modules as backbone and exploits DNNs to learn the compositionality of the human body by enforcing a bone-based part representation as the output of intermediate stacks. With the use of only five hourglass modules, it outperforms other methods while being computationally cheaper. A more detailed description of all the used architectures is provided in the supp. mat. Sec. 1.

Whenever possible we use the released networks from the authors, otherwise we implement ourselves. Further, we use a standard protocol to evaluate the performance for different networks on the validation sets that includes similar cropping and data pre-processing. Therefore, our reported results might be a little inferior to the reported results that employ flipping, multiple crops and other similar techniques. In order to show the generalizability of our findings, we study two different pose databases - MPII [1] and COCO [21]. We use PCKh [1] and OKS [17] as metrics for MPII and MS COCO, respectively. All the results are reported on the validation set. Due to space constraint, we show the results on MPII in this manuscript and refer to the supp. mat. for the results on MS COCO.

3.1.1 Adversarial Attack Methods

Theoretically, adversarial attack consists of adding an adversarial noise n , to the input I , of a network $f(x; \theta)$, that changes the output $y = f(I; \theta)$.

Fast Gradient Sign Method [14] which explicitly bound the maximum magnitude (l_∞ norm) of every pixel are most popular and relatively computationally cheap. FGSMs use the scaled, by ϵ , sign of gradient w.r.t. the desired objective to obtain $n : \|n\|_\infty < \epsilon$. They can either be targeted or untargeted and single-step or iterative. An untargeted FGSM attack (**FGSM-U**) simply increases the loss of the network for a given input I to obtain perturbed input I^p as-

$$I^p = I + \epsilon \cdot \text{sign}(\nabla_I \mathcal{L}(f(I; \theta), y)) \quad (1)$$

Whereas, a targeted FGSM attack (**FGSM-T**), pushes the output of the network towards a target y^t . For classification systems, y^t can be easily obtained as the least likely or target output of the network [20]. Unfortunately, HPE systems do not have a least likely target pose for a given input image. Therefore, we choose at random one target pose, P^t from a pool of ground-truth poses from the validation set, $\mathcal{P} = \{\hat{P}_1, \hat{P}_2, \dots\}$, that gives a PCKh value of 0 for the

predicted $P = f(I; \theta)$. This can be construed as selected the most unlikely pose for a given image and leads to-

$$I^p = I - \epsilon \cdot \text{sign}(\nabla_I \mathcal{L}(f(I; \theta), P^t)) \quad (2)$$

Both untargeted and targeted FGSM attacks, can be extended to their iterative counterparts **IGSM-U-N** and **IGSM-T-N**, respectively, that iterate N times to yield the final perturbed image I^p starting with I . The perturbed image I_i^p for the i^{th} iteration for untargeted (Eq. 3a) and targeted (Eq. 3b) attack is given as-

$$I_i^p = C_\epsilon(I, I_{i-1}^p + \alpha \cdot \text{sign}(\nabla_{I_{i-1}^p} \mathcal{L}(f(I_{i-1}^p; \theta), y))) \quad (3a)$$

$$I_i^p = C_\epsilon(I, I_{i-1}^p - \alpha \cdot \text{sign}(\nabla_{I_{i-1}^p} \mathcal{L}(f(I_{i-1}^p; \theta), P^t))) \quad (3b)$$

$$s.t. x_0 - \epsilon \leq C_\epsilon(x_0, x_i) \leq x_0 + \epsilon \quad (3c)$$

where, $C_\epsilon(x)$ clips x to $[x - \epsilon, x + \epsilon]$.

All the aforementioned attacks are image-specific and require costly back-propagation through the network for its computation. Therefore, [24] proposed to learn image-agnostic or *universal perturbations* from a representative subset of images for a given image distribution. In our experiments, however, we adopt the method in [15] to HPE systems and obtain the universal perturbation u . Its an iterative process that computes perturbations on training samples x_i , or mini-batches of them, and aggregates them to obtain the final u after re-scaling-

$$u = u + \delta \cdot \text{sign}(\nabla_{x_i} \mathcal{L}(f(x_i; \theta), y)) \quad (4)$$

We fix $\delta = \frac{\epsilon}{200}$, mini-batch size of 16 and $\|u\|_\infty \in \{8, 16\}$, because lower ϵ values hindered learning while higher values are perceptible and use the same setting for all the architectures. The obtained u can be simply added to any image to attack the network, therefore, making it more widely applicable than network access attacks.

Since the performance of the used models differ, it is not fair to compare the degradation due to adversarial attacks using the drop in absolute performance. Therefore, for untargeted and universal attacks, we report $(\text{perturbed}/\text{original}) * 100$ score ratio for which lower values indicate more effective attack. For the targeted attacks, we report the target PCKh score of the output w.r.t. to the target, therefore, higher values indicate more effective attacks. The degree of intensity which measured by the maximum permissible pixel differences between I^p and I and denoted by ϵ is varied in $\{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$. For iterative attacks, we have chosen to report the effects under a setting similar to that popularly employed to attack classification systems and limit the maximum number of iterations to 10, but the HPE systems are relatively robust, therefore, we also report the result with a maximum of 100 iterations. However, the targeted attacks are still difficult, therefore, they require 20 iterations. Overall, it yields

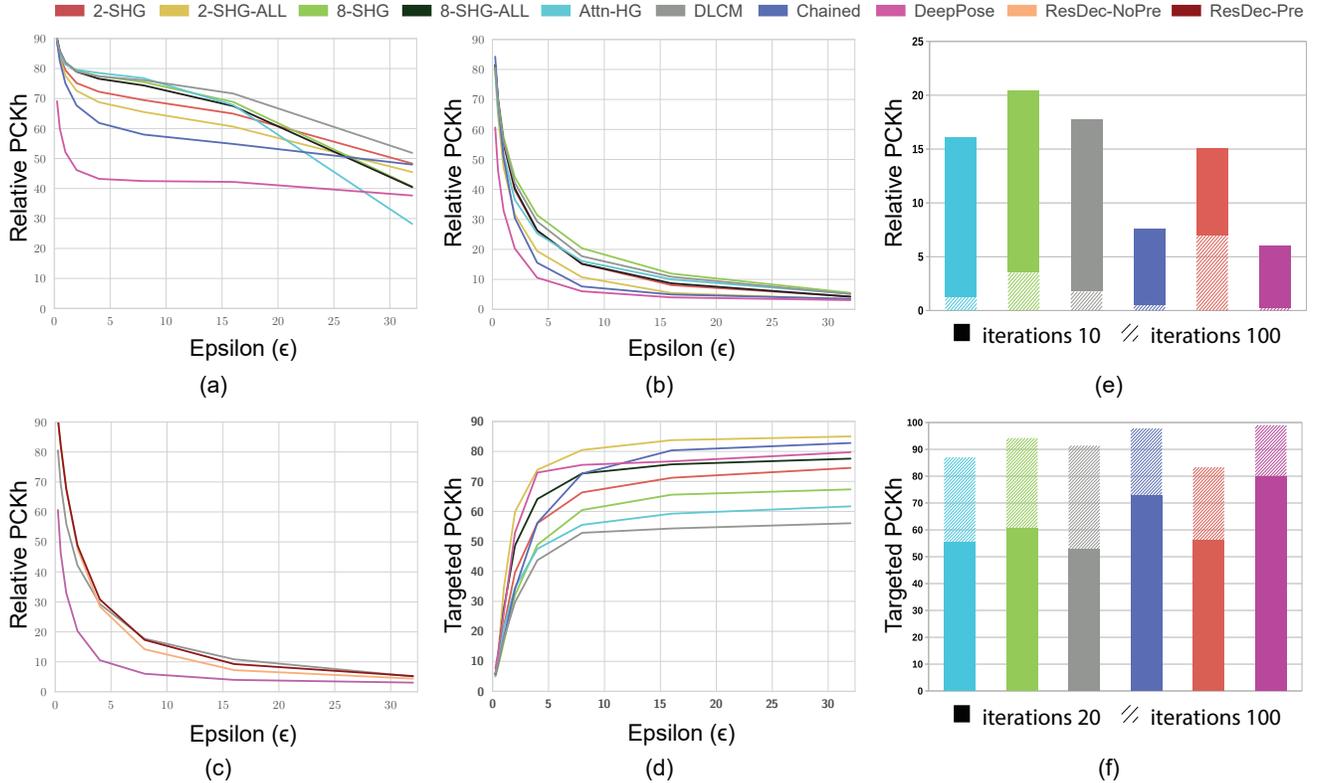


Figure 2. Comparison of different types of attacks on all the models. (a) and (b) depict the relative PCKh as a function of ϵ for FGSM-U and IGSM-U-10, respectively. (c) emphasizes the difference between direct regression and heatmap under IGSM-U-10 attack. (d) depicts the final PCKh with respect to the target for IGSM-T-20. (e) & (f) shows the relation between number of iterations and effectiveness of untargeted and targeted attacks, respectively.

four different configurations of attacks **IGSM-U/T-10/20**, and **IGSM-U/T-100/100**. Also, we observe that the optimal value of the step-size, α , falls in the range $[\frac{\epsilon}{3}, \frac{\epsilon}{2}]$ for untargeted and in $[\frac{\epsilon}{9}, \frac{\epsilon}{7}]$ for targeted attacks. We report the results of IGSM-U/T-100 with $\epsilon = 8$ and refer to supp. mat. Sec. 4 for other values of ϵ , while IGSM-U/T-10/20 results are reported for all ϵ values. Since this is a preliminary work on attacks on HPE, we stick to the standard attack mechanisms to provide insights into the problem.

4. Adversarial attack on HPE systems

This section starts with White Box Attacks, where we have complete access to the target network, and study its effect under varying ϵ , number of iterations, architectures and targeted vs. untargeted setting. Then we report results on image-agnostic universal perturbations with varying ϵ and different architectures with their visualizations to shed light on their workings. We also report the effect of both attacks in *black-box* mode, in which we learn the perturbation using one network and use it to attack a target network to which we have no access. We also report the vulnerability of different body joints towards adversarial attack followed by a discussion of interesting insights pertaining the different

architecture’s robustness and effect of some simple image processing based defense strategies. We also performed a similar study of COCO [21] dataset and can be found in supp. mat. Sec. 5.

4.1. White Box Attacks

The complete access to a network exposes it to a variety of different attacks. The main result for this section is shown in Fig. 2 that plots the effect of FGSM-U, IGSM-U-10 and IGSM-T-20 attacks vs. ϵ on different HPE architectures described in Sec. 3.1.

4.1.1 HPE vs. Classification Systems

We first compare the robustness of HPE systems in general to another task that involves per-pixel reasoning, semantic segmentation (presented in [2]). A simple comparison between the relative drop in the performance for FGSM-U attack on HPE Fig. 2a and semantic segmentation (ref. [2] Fig. 2(a)) reveals that the HPE systems undergo less degradation. While some part of the observed relative robustness can be attributed to a more lenient metric, PCKh vs. IoU. We believe that some of it perhaps comes from the successive down-sampling and up-sampling of the HourGlass in-

roduces multi-scale processing, which has been previously reported to be effective against adversarial attacks on semantic segmentation

4.1.2 Robustness of Different Models

The observations from Fig. 2 reveal that the order of robustness of different models across different attacks is more or less consistent. We can observe that the heatmap based approaches are more robust than direct regression (DeepPose) based approach. This is because the direct-regression loss function is also a measure of PCKh after thresholding while heatmap loss produces Gaussian bumps at joint-location, which is not as strongly correlated to PCKh. Also, heatmap predictions, unlike regressed values, are implicitly bounded to be valid image coordinates.

In order to make a fair comparison between, we use the same ResNet backbone and use a simple regression loss in one case, and de-conv layers followed by heatmap regression in the other case. We name them as **ResDec-Pre** and **ResDec-NoPre** for resnet-deconvolution with and without imagenet pretraining. As seen in Fig 2c, relative performance for untargeted attacks is noticeably higher for heatmap loss. Also for ResDec-Pre, the relative performance is even higher, validating the findings of [16]. Strikingly, ResDec-Pre is almost as robust as the most robust network - DLCM. This advocates a requirement to move away from the popular regression-based 3D-HPE frameworks [10, 23, 32, 44] (see supp. mat. Sec. 8 for details on 3D-HPE experiments). We leave theoretical understanding of robustness caused by imagenet pretraining a question for future study.

Due to the conditional joint prediction nature of the architecture that propagates the perturbation in one joint to the rest of the joints, Chained-Prediction turns out to be the least robust among the heatmap-based approaches. We observe that DLCM is more robust than 2/8-SHG and AttnHG against all attacks, perhaps due to DLCM’s imposition of human skeleton topology. This encourages further exploration of structure-aware models to counter adversarial examples.

4.1.3 Effect of the Number of Iterations on the Attack

Fig. 2e 2f plots the relative drop and target PCKh for untargeted and targeted attacks, respectively, for $\epsilon = 8$ with 10 and 100 iterations. We observe that moving from 10 to 100 iterations results in dramatic degradation for all the networks under both the settings. This observation is in contrast with the effect of IGSMs on classification or semantic segmentation problems, where [19] finds that $\min(\lceil 1.25\epsilon \rceil, \epsilon + 4)$ iterations are sufficient for complete degradation. HPE, on the other hand, often needs up to 100

iterations for the same. Unfortunately, with enough iterations, all the systems degrade by over 95% which shows that all models are vulnerable for carefully designed perturbations. See supp. mat. Sec. 4 for results on all ϵ values.

4.1.4 Stacked Hourglass Study

Since most HPE systems build on the Stacked-Hourglass backbone [26], we carry out a thorough analysis of adversarial attack on SHG architecture with different network hyper-parameters such as depth (number of stacks). First, we find that increasing the number of hourglasses from 2 to 8 increases the robustness of the model; a fact clearly visible from Fig. 2a 2b 2d. Next, we study the effect of simultaneous perturbation of outputs of all the stacks of SHG, indicated by suffix *ALL*, and observe that the attacks become more effective, again evident from Fig. 2a 2b 2d. Specifically, 2-SHG-ALL and 8-SHG-ALL attacks increased the target PCKh from 66.3 to 80.5 and from 60.5 to 73.0, respectively. This is expected because downstream stacks are supposed to improve upon the predictions of the upstream ones and hence, incorrect prediction upstream will cascade into errors in the final output. Further, intermediate supervision would provide better gradient flow especially since the stacks are not connected via residual connections. Interestingly, 2-SHG-ALL IGSM-T-20 attack brings down its performance even below Chained-Prediction and DeepPose in, the two worst performing architectures in terms of robustness to adversarial attacks!

4.1.5 Targeted vs. Untargeted Attacks

Targeted attacks are more difficult than untargeted ones as evidenced from the fact that targeted attacks require higher number of iterations as compared to an untargeted attack, 20 vs. 10. It is because an untargeted attack can simply take large steps in the direction of increasing loss for I , whereas, the targeted attack requires finding the optimal $I^p : \|I - I^p\|_\infty \leq \epsilon$ where the loss $\mathcal{L}(f(I^p; \theta), P^t)$ is small; a more difficult problem. We observe that the optimal value of step-size α for IGSM-T is found to be almost 3 times smaller than that of IGSM-U as expected. However, small step-size based iterative targeted attacks with sufficient iterations, around 100, can still lead to almost 100% target PCKh Fig 2e, 2f. As ϵ increases, different architectures under untargeted attack converge in performance while they diverge for targeted attacks! It indicates that under extreme targeted attack different networks perform significantly different in terms of their robustness. It is worth noting that the Relative PCKh (relative degradation w.r.t. original target) was almost equal in both IGSM-U-10 & IGSM-T-20 (refer to supp. mat. Tables 4,6).

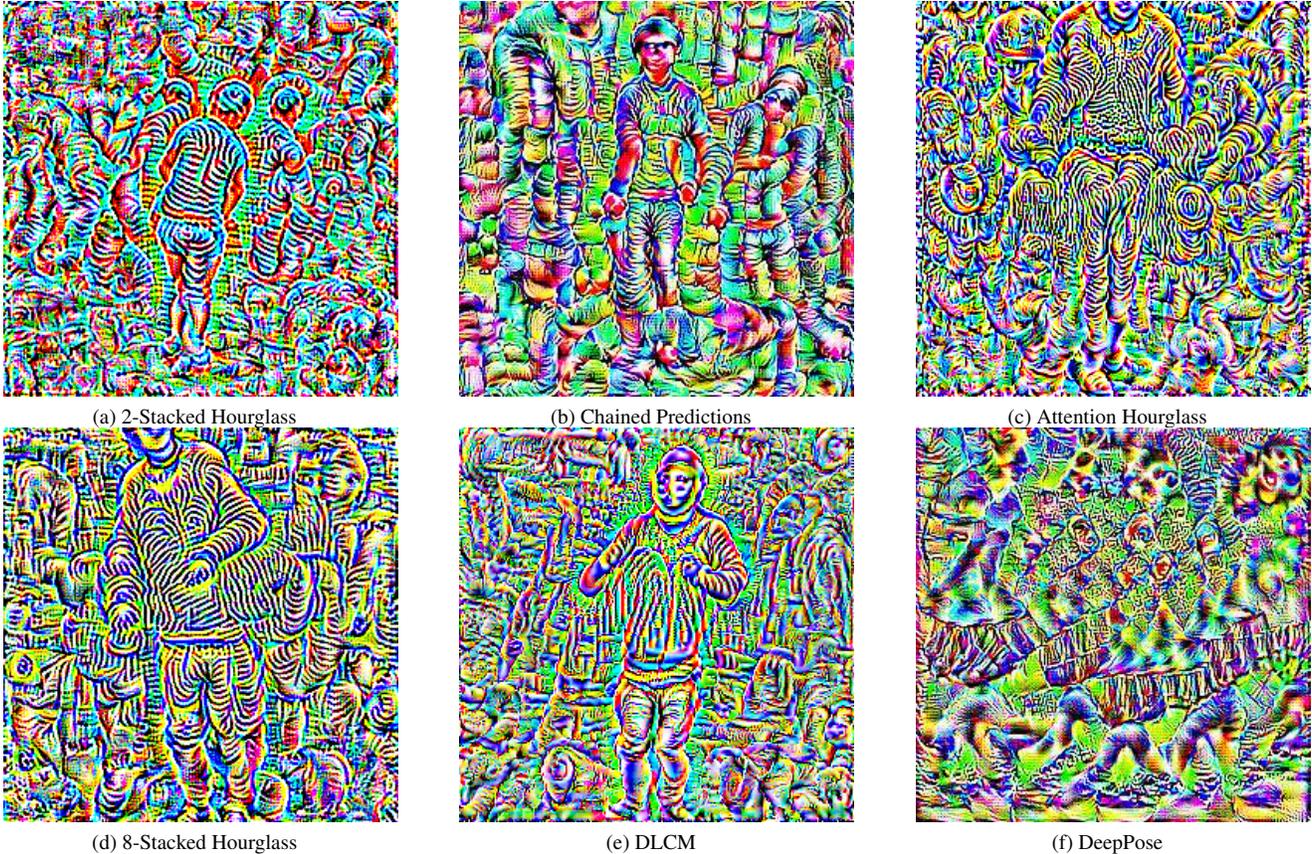


Figure 3. Visualization of image-agnostic universal perturbations, with $\epsilon = 8$, for different networks scaled between 0 to 255 for better visualization. Note the hallucinated body-joints, mostly arms and limbs to fool HPE networks. More vis. in supp. mat. Fig 4-6

| | | 8-SHG | 8-SHG-ALL | Attn-HG | DLCM | 2-SHG-ALL | 2-SHG | Chained | DeepPose | Vulner. |
|----------------|----------|--------------|-------------|--------------|--------------|--------------|-------------|--------------|-------------|---------|
| Target Network | 8-SHG | <u>8.85</u> | <u>5.92</u> | 53.32 | 56.61 | 53.45 | 68.17 | 63.23 | 86.7 | 63.58 |
| | Attn-HG | 41.92 | 48.47 | <u>11.47</u> | 57.62 | 61.05 | 71.68 | 68.1 | 84.78 | 61.95 |
| | DLCM | 46.76 | 47.09 | 60.07 | <u>12.75</u> | 64.45 | 74.02 | 67.41 | 84.93 | 63.53 |
| | 2-SHG | 51.95 | 55.17 | 75.28 | 70.08 | <u>10.35</u> | <u>15.7</u> | 51.6 | 88.59 | 65.45 |
| | Chained | 77.65 | 79.7 | 82.57 | 81.15 | 72.08 | 78.45 | <u>10.96</u> | 75.36 | 78.14 |
| | DeepPose | 74.19 | 70.44 | 75.12 | 75.03 | 72.23 | 75.6 | 42.04 | <u>2.78</u> | 69.24 |

Table 1. The results of all source-target pairs under doubly black-box attack setting. Rows represent the relative degradation in the target network when attacked by the network in the column. **Vulner.** stands for ease of attack under doubly black-box setting. **Boldface** shows the strongest black box attack for a model and underlined numbers indicate the performance of the model on itself

4.2. Image-Agnostic Adversarial Perturbations

We follow Sec. 3.1.1 to obtain the universal adversarial perturbations for all the considered architectures. Once obtained, they can be simply added to any input image to fool the corresponding architecture, making them practically useful in real-world scenario. Fig. 3 shows the universal perturbations, scaled between 0 to 255 for better visualization (more visualizations can be found in supp. mat. Fig 4-6). It is, to the best of our knowledge, the first visualization of such perturbations for HPE, which reveal semantic hallucinations. A closer look reveals that universal perturbations confuse HPE systems by hallucinating body-joints,

mostly limbs, throughout the image. Visual inspection of the skeletons predicted on these perturbations reveal similarity with hallucinated joints and can be found in Supp. mat. Fig. 8-13. Even more surprisingly, some networks have similar prediction across different images despite the fact that these perturbations were not explicitly designed to predict these specific outputs. It is worth noting that while all visualizations of UAP resemble the human body, visualization of DeepPose UAP does not do so and since the UAP are computed as the gradient averaged over all training images, this means that the heatmap based approaches have minimized loss when the joints are discernible, but DeepPose has not.

| Model and Attack | Ankle | Knee | Hip | Neck | Head | Shoulder | Elbow | Wrist |
|------------------------|--------------|-------------|--------------|--------------|--------------|----------|-------|--------------|
| Relative PCKh | | | | | | | | |
| DeepPose-UI | 0.63 | 1.24 | 4.43 | <u>17.52</u> | 13.11 | 4.39 | 2.35 | 2.2 |
| 2-SHGlass-UI | 3.82 | 4.62 | 2.82 | <u>41.89</u> | 23.39 | 24.79 | 14.48 | 13.4 |
| 8-SHGlass-UI | 8.79 | 10.9 | 3.04 | <u>45.54</u> | 34.61 | 29.67 | 20.65 | 20.54 |
| Chained-Predictions-UI | 2.79 | 2.07 | 3.53 | <u>22.7</u> | 15.73 | 11.87 | 4.05 | 3.77 |
| Attention-HG-UI | 6.52 | 7.61 | 3.05 | <u>39.31</u> | 25.01 | 21.35 | 17.54 | 16.96 |
| DLCM-UI | 6.28 | 6.79 | 2.12 | <u>45.69</u> | 29.72 | 28.04 | 17.75 | 16.4 |
| <i>Average</i> | 4.81 | 5.54 | 3.12 | <u>35.44</u> | 23.60 | 20.02 | 12.80 | 12.22 |
| Target PCKh | | | | | | | | |
| DeepPose-TI | <u>59.22</u> | 73.04 | 84.79 | 81.64 | 73.0 | 82.4 | 77.93 | 69.33 |
| 2-SHGlass-TI | 62.61 | 69.65 | 86.0 | 70.05 | <u>49.79</u> | 72.73 | 64.68 | 47.59 |
| 8-SHGlass-TI | 48.24 | 54.02 | 83.86 | 71.94 | 51.38 | 70.53 | 56.33 | <u>43.55</u> |
| Chained-Predictions-TI | 70.9 | 77.53 | 84.59 | 74.64 | <u>59.29</u> | 75.26 | 72.28 | 60.2 |
| Attention-HG-TI | 47.25 | 52.06 | 77.97 | 60.46 | <u>39.53</u> | 57.22 | 52.59 | 48.62 |
| DLCM-TI | 47.8 | 54.99 | 74.63 | 57.93 | <u>38.88</u> | 55.26 | 48.58 | 39.57 |
| <i>Average</i> | 56.00 | 63.55 | 81.97 | 69.44 | <u>51.97</u> | 68.9 | 62.07 | 51.47 |

Table 2. Relative PCKh of different body-joints for untargeted attacks across different networks. **Boldface** and underlined numbers indicate the most and the least vulnerable joints, respectively. Note that hips, knee and ankles are more vulnerable than the rest.

Universal perturbations degrade the original performance, averaged over all models, on the training (used to obtain them in the first place) and validation sets to 6.4% and 9.9% of their original value, respectively with $\epsilon = 16$. It clearly showing their strong effect, see supp. mat. for results with $\epsilon = 8$. Network-wise results on the effect of universal perturbations are reported in Table 1. Surprisingly, their effect on the performance is similar in magnitude to image-specific iterative attacks, 9.9% vs. about 8% for latter ($\epsilon = 16$). So these are computationally efficient while being equivalent to Image-Dependent methods. We also study the dependency of universal perturbations on the amount of training data needed, as in [24], by obtaining them with varying number of samples from the training set. Please refer to supp. mat. Sec. 4.4 that shows the variation of degradation ratio vs. number of samples. We observe that even with 10% data samples, i.e. only 2500 images, the obtained universal perturbations degrade the performance to 18% vs. 9.9% with all the 25925 samples.

4.3. Black-Box Attacks

This setting refers to an attack on *target* network using adversarial perturbations learned from a different network, referred to as *source* network. We do not have access to the target network at any stage except while evaluating the performance. The perturbations can either be image-specific, obtained by FGSM-U/T or IGSM-U/T from the input image, or image-agnostic universal perturbations. The latter gives rise to *doubly black-box* attacks i.e. we need neither access to the target network nor do we use the image to obtain the perturbation. We report all the combinations of ($S \rightarrow T$) pairs and tabulate the results in Table 4.1.5 in supp. mat., due to space constraints. In general, we observe

30-40% degradation in the target network’s performance.

Doubly black-box attacks are reported in Table 1 where we can again observe fair generalization with 30-40% cross-network degradation, on an average. We observe that the generalization is stronger across similar architectures. Specifically, Stacked-Hourglass’s perturbation degrades DLCM and Attention-Hourglass to 50%, but DeepPose and Chained-Prediction to only 75%.

4.4. Body-Joint Vulnerability Towards Attack

In order to understand the effect of adversarial attack on different body joints, we report per-joint accuracy under different architectures and attack-types for MPII dataset in Table 2. For left-right symmetric body-joints (ankle, knee, hip, shoulder, elbow and wrist), we report the left-right average degradation. Its evident that head and neck are the most robust while hips are the most vulnerable across different attacks. It could be due to the fact that the HPE networks are trained on cropped images that have tightly localized head in most of the samples, whereas limbs are spread throughout the images at weird locations. Therefore, it is difficult to fool the network in predicting head and neck in some other region. Moreover, we observe that the relative performance of different joints vary dramatically for untargeted attacks while it doesn’t vary so much for targeted attacks. These observations can motivate future work focus on understanding and improving robustness of the more vulnerable joints.

5. Simple Image Processing for Defense

In this section we discuss the effect of simple image-processing based defense strategies against adversarial attacks on HPE systems. Since this is a preliminary work on

adversarial attacks on human pose, we focus only on computationally cheap methods to mitigate the effect of the different attacks.

Recently [40], showed that the adversarial attacks in semantic segmentation can be detected by analyzing the consistency of the predicted segmentation map. Similar reasoning can be extended to HPE systems and we thought that the predicted skeletons from adversarially perturbed image would look unrealistic. Surprisingly, visual inspection of the skeletons reveals that the skeletons are semantically meaningful. It could be due to the implicit learning of human-body structure that prevents the networks from producing structurally garbage results even after adversarial attacks. Secondly, we thought of checking the quality of Gaussian bumps under adversarial attack thinking that they might distort from being Gaussian. Again, we observe that the bumps still resemble Gaussian which can be quantitatively measured using the KL divergence and is reported in the sup. mat. Sec. 2. Therefore, even this measure cannot be used for detecting the presence of adversarial attack.

We also tried simple geometric and image-processing based defense strategies like flipping and smoothing. As expected, smoothing worked well for both image-specific and image-agnostic attacks, a finding supported by multiple research work in the past [2, 31]. Also, we observe that flipping an image-specific perturbations renders it relatively ineffective. Specifically, a non-flipped version of image-specific perturbation degrades the network to a range of 5-10% whereas, its flipped version can only reduce it to about 70-75%. This shows that image-specific perturbations are *truly specific* and don't work with flipping. On the other hand, universal perturbations were equally detrimental under flipping too! It can easily be explained on the basis of the fact that universal perturbations are generic while image-dependent perturbation are very specifically aligned. The same is also evident from the visualization of universal perturbations.

6. Conclusion

We performed a dense and exhaustive analysis of various adversarial attacks on human pose estimation systems, using MPII [1] & COCO [21] and found some interesting trends in how design choices affect robustness. We report that the image-agnostic universal perturbations are as detrimental an attack as image-specific iterative approaches while being computationally much cheaper to obtain. Our visualizations of universal perturbations exhibit a strikingly human-like hallucinated array of body-joints to fool the networks. Further our analyses on the vulnerability of different joints helped identifying the most and least robust body parts under adversarial attack.

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 3, 8
- [2] Anurag Arnab, Ondrej Miksik, and Philip H.S. Torr. On the robustness of semantic segmentation models to adversarial attacks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 4, 8
- [3] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 2687–2695, 2018. 2
- [4] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2613–2621. Curran Associates, Inc., 2016. 1, 2
- [5] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012. 2
- [6] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57. IEEE Computer Society, 2017. 1, 2
- [7] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Robust physical adversarial attack on faster R-CNN object detector. *CoRR*, abs/1804.05810, 2018. 1, 2
- [8] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 3
- [9] Moustapha M Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6977–6987. Curran Associates, Inc., 2017. 2
- [10] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, 2018. 5
- [11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2
- [12] Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. Adversarial examples for semantic image segmentation. *CoRR*, abs/1703.01101, 2017. 1, 2

- [13] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. 2016. 2, 3
- [14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. 2, 3
- [15] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3
- [16] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *arXiv*, 2019. 5
- [17] <http://cocodataset.org/keypoints eval>. OKS Metric for keypoint detection evaluation. Technical report. 3
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 2
- [19] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. 2, 5
- [20] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *CoRR*, abs/1611.01236, 2016. 1, 2, 3
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1, 3, 4, 8
- [22] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *CoRR*, abs/1611.02770, 2016. 1, 2
- [23] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 5
- [24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3, 7
- [25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2
- [26] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 483–499, 2016. 1, 2, 5
- [27] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2
- [28] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016. 2
- [29] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387, 2016. 1, 2
- [30] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2
- [31] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James A. Storer. Deflecting adversarial attacks with pixel deflection. *CoRR*, abs/1801.08926, 2018. 8
- [32] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 5
- [33] Sayantan Sarkar, Ankan Bansal, Upal Mahbub, and Rama Chellappa. UPSET and ANGRI: Breaking high performance image classifiers. *CoRR*, abs/1707.01159, 2017. 2
- [34] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earleence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *WOOT @ USENIX Security Symposium*. USENIX Association, 2018. 1, 2
- [35] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? – a comprehensive study on the robustness of 18 deep image classification models. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2
- [36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. 2
- [37] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3
- [38] Jonathan J. Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1799–1807, 2014. 2
- [39] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1653–1660, 2014. 2
- [40] Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, and Dawn Song. Characterizing adversarial examples based on spatial consistency information for semantic seg-

- mentation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 8
- [41] Chaowei Xiao, Bo Li, Jun yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3905–3911. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 2
- [42] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2
- [43] Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. Fooling vision and language models despite localization and attention mechanism. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2
- [44] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, 2017. 5