Improving the Robustness of 3D Human Pose Estimation: A Benchmark Dataset and Learning from Noisy Input

Trung-Hieu Hoang[1] Mona Zehni [1] Huy Phan[2] Duc Minh Vo[3] Minh N. Do[1]
[1]University of Illinois at Urbana-Champaign, USA
[2]VinUniversity, Ha Noi, Vietnam [3]The University of Tokyo, Japan
{hthieu,mzehni2,minhdo}@illinois.edu, 20huy.pn@vinuni.edu.vn, vmduc@nlab.ci.i.u-tokyo.ac.jp

Abstract

Despite the promising performance of current 3D human pose estimation techniques, understanding and enhancing their robustness on challenging in-the-wild videos remain an open problem. In this work, we focus on building robust 2D-to-3D pose lifters. To this end, we develop two benchmark datasets, namely Human3.6M-C and HumanEva-I-C, to examine the resilience of video-based 3D pose lifters to a wide range of common video corruptions including temporary occlusion, motion blur, and pixel-level noise. We demonstrate the poor generalization of state-of-the-art 3D pose lifters in the presence of corruption and establish two techniques to tackle this issue. First, we introduce Temporal Additive Gaussian Noise (TAGN) as a simple yet effective 2D input pose data augmentation. Additionally, to incorporate the confidence scores output by the 2D pose detectors, we design a confidence-aware convolution (CA-Conv) block. Extensively tested on corrupted videos, the proposed strategies consistently boost the robustness of 3D pose lifters and serve as new baselines for future research.

1. Introduction

Human pose estimation in 3D (3D HPE) from a monocular RGB video is a challenging computer vision task with a wide range of applications in action recognition [21, 22], virtual/augmented reality [1], human-computer interaction [54] and healthcare [11], to name a few. In this paper, we focus on the robustness of 2D-to-3D HPE that utilize *off-the-shelf 2D keypoint detection* followed by a *2D-to-3D pose lifter* to lift the sequence of detected 2D keypoints to 3D camera space [32, 60, 64]. While promising results have been shown on standard benchmarks [13, 24, 42], with minimal subjects occlusion, the real-world recordings are far from this controlled setting due to the appearance of external objects or improper camera's field of view [36, 38, 53]. We made an important observation that *these models are less robust to occlusions and disruptions in video appearance*. As this is a two-stage process, one might suggest improving the robustness of the 2D pose detectors [14] instead of the 3D lifter. While this is a valid solution, it is not practical in some real-world applications, such as healthcare, where it is computationally expensive or violates patient's privacy when operating on RGB videos [11, 43]. Hence, this work benefits systems that favor using off-the-shelf 2D pose estimator, operating on a sequence of detected keypoints without touching the RGB videos.

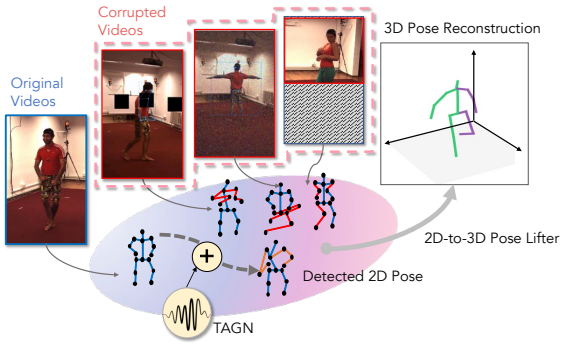

Figure 1. Illustration of a 2D-to-3D pose lifter operating on 2D poses detected from corrupted video frames (red boxes). Our proposed Temporal Additive Gaussian Noise (TAGN) serves as a 2D pose augmentation that adds jitter to the detected 2D poses of the original video frames (blue box). TAGN's goal is to improve generalization on test videos with unforeseen visual corruptions.

We first systematically study the robustness of 3D HPE by *providing new benchmarks*, namely Human3.6M-C and HumanEva-I-C based on [13, 42]. Following [53], we refer to *corruption* as image-level visual distortions (such as noise, blur, or pixelation). We define frame- and video-level corruption operators inspired by real-world conditions (e.g., partial occlusions or random noises), and use them to construct visually corrupted video datasets. These datasets are obtained after augmenting videos in [13, 42] with corruptions such as cropping or geometric occlusions. We demonstrate the poor robustness of these models when the test time 2D keypoints are erroneous.

Furthermore, *we introduce two baseline solutions* to en-

hance the robustness of 3D pose lifters under visual corruption problems in *two scenarios* when the corruptions at test time are known and unknown. Firstly, we introduce a new *data augmentation* strategy with temporal additive Gaussian noise (TAGN), which can be plug-and-play to any existing methods. TAGN randomly perturbs the 2D keypoints during training, to further simulate 2D input uncertainty and improve the robustness of the pose lifting models at test time. We illustrate the proposed approach in Figure 1. Secondly, confidence-aware convolution (CA-Conv), an *elaborated version* of the regular temporal convolution block is proposed. While many 2D human pose estimators output a confidence score for each detected keypoint, this information is often discarded when passed to the 3D pose lifter. CA-Conv properly combines the 2D keypoints and their confidence scores to achieve a robust lifting task. We summarize our main contributions as below:

- We propose the first *synthesized video-based 3D HPE datasets* (Human3.6M-C and HumanEva-I-C) explicitly designed to evaluate the robustness of 3D pose lifters. We also introduce a proper MPJPE-based evaluation metric to quantify the robustness and extensively analyze the performance of several state-of-the-art 2D-to-3D pose lifting models [21, 32, 40, 60, 65].
- We develop TAGN, a simple yet *effective augmentation strategy* that can be conveniently applied while training any 2D-to-3D pose lifter. Unlike image-based augmentations, TAGN is applied to the input 2D keypoints, leading to better test-time generalization than models trained on uncorrupted datasets.
- We present CA-Conv, a *confidence-aware extension* of the regular temporal convolution block, and show its efficacy when trained on corrupted videos. To the best of our knowledge, this is the first 3D pose lifter consuming the confidence score associated with the input 2D pose.

2. Related Works

3D Human Pose Estimation: We focus our review on methods devoted to single-person 3D HPE tasks from monocular RGB images and videos. This problem is inherently complex due to depth ambiguity and occlusions. We study 3D HPE under two broad categories. One line of work learns to directly estimate the 3D pose in an end-to-end manner from the 2D RGB image with no intermediate supervision [17, 25, 29–31, 41, 45, 48–50, 66, 67]. In the second class of 3D HPE methods, 2D poses (i.e., detected 2D keypoints of human body joints) are firstly estimated given RGB images [56, 63] and subsequently lifted to 3D. While direct regression-based 3D HPE methods have degraded performance compared to 2D-to-3D lifting due to the lack of intermediate supervision, the main challenge for the lifter models is associated with the inaccuracies of the detected 2D pose [23, 32]. Among pioneering works, Chen *et al.* [4] approached the 2D-to-3D pose lifting problem as nearest neighbor matching from large 3D mocap datasets. Martinez *et al.* [23] trained a deep network with linear layers to predict 3D pose from estimated 2D pose. Other 2D-to-3D pose lifting approaches estimate distance matrices [28], employ evolutionary and differentiable data augmentations [9, 18], or devise self-supervised methods [51]. For 3D pose lifting in videos, the sequential nature and the temporal dependence of the frames are taken into account by processing a temporal sequence of the 2D pose. Temporal dilated convolutions [32], graph convolution networks [12, 55, 61] and transformer-based models with temporal and spatial attentions [14, 19, 65] are among the current video-based 3D pose lifting solutions.

Robust 3D Human Pose Estimation: Robustness to noise and occlusion is a desired property for any real-world ready 3D HPE model. Recent occlusion-aware or noise-resilient 3D HPE solutions are either semantically occlusion-aware depth-based methods [34] or well-designed convolutional models with occlusion-aware heatmaps [16, 62]. However, these methods operate on single images or depth maps and are not designed for a video setting. 3D human pose modeling with partial visibility especially for consumer videos [3, 37] and depth-aware techniques for multi-person pose estimation [46, 47], are among video-based HPE solutions tackling the robustness aspect. Also, [7] addressed scenarios with self-occlusion by introducing a cylinder man model for pose regularization and occlusion augmentation while adopting optical flow on 2D keypoint heatmaps. [6] considered various frame- or keypoint-level occlusion augmentations on 2D heatmaps. Furthermore, [36] builds a motion prior and solves a test-time optimization to find plausible 3D poses coherent with the 2D non-occluded keypoints. Although effective, these methods can be computationally expensive and unsuitable for real-time applications. In addition, [40] incorporated self-supervised pre-trained denoising auto-encoders in a 2D-to-3D human pose estimator. However, the improvement in robustness on corrupted 2D keypoints has not been explicitly shown.

Learning with Additive Jitter: Training with jittered input is a simple regularization strategy in machine learning that has been shown to improve generalization [35]. By evading overfitting, it also enhances the robustness and accuracy of neural networks [8]. Besides, in the field of interpretable machine learning, it has proven helpful in raising the awareness of models' explanations of the input uncertainty [52].

Corrupted Dataset for 3D HPE: Corrupted datasets are widely used to analyze the robustness of machine learning models in many tasks such as image classification [10], object detection [27], semantic segmentation [15], or 2D HPE [53]. Meanwhile, occlusion is an HPE domain-specific challenge. Sárándi *et al.* in [38] introduced geometric occlusions in various shapes such as circles, rectangles,

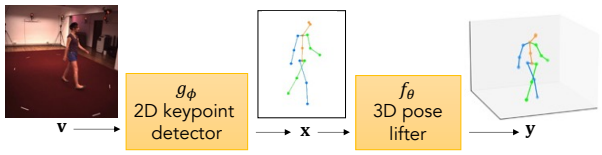

Figure 2. 3D human pose estimation in a 2D-to-3D pose lifting pipeline. The detected 2D pose by g_ϕ is lifted to 3D by f_θ.

or bars on Human3.6M dataset [13] and showed these augmentations improve test-time performance. More realistic copy-paste augmentations to synthesize occlusion were developed in [2, 39]. However, to the best of our knowledge, no existing work studies the robustness of lifter-based 3D HPE models on synthetically corrupted videos.

Our Work: Compared to prior works with single-frame synthesized occlusions [38, 39, 53], our benchmark is more comprehensive by including both complex and realistic disruptions like temporal occlusion or motion blur that better match the temporal nature of video data. The closest to our corrupted dataset is [53], devoted to the single image 2D HPE task. Meanwhile, we focus on the robustness of *temporal* 3D pose lifters [32, 65] with noisy 2D input keypoints from a corrupted video. This enables us to better study temporal occlusions and their impact. Furthermore, we provide solutions to enhance robustness to errors in 2D pose input, by learning with additive jitter, and taking into account the confidence score of the detected 2D poses when lifting.

3. Notations and Problem Setup

Notations: Consider a two-stage 3D HPE solution as illustrated in Figure 2. Let g_ϕ and f_θ (parameterized by ϕ and θ) denote the 2D keypoint detection (such as OpenPose [63] or HRNet [44]) and 3D pose lifter modules, respectively. To take into account the temporal nature of the video data, we consider lifter models that consume a sequence of T 2D poses. A sequence of RGB frames (of $H \times W$ resolution) $\mathbf{v} \in \mathbb{R}^{T\times 3\times H\times W}$ is passed as input to g_ϕ. Next, the 2D detected pose with J keypoints, i.e. $\mathbf{x} = g_\phi(\mathbf{v})$, $\mathbf{x} \in \mathbb{R}^{T\times J\times 2}$, is processed by f_θ to finally output the estimated 3D pose for the middle frame of the input sequence $\mathbf{y} = f_\theta(\mathbf{x})$, $\mathbf{y} \in \mathbb{R}^{J\times 3}$. Padding is appropriately applied either at the start or end of the sequence. We *fix* g_ϕ to be an off-the-shelf 2D HPE model that is already trained.

Robustness of 3D pose lifters: We interest in the effect on f_θ where the *input RGB video frames are randomly corrupted* by a set of image-domain operators (the specific set of operators will be introduced in Sec. 6.1), simulating undesirable circumstances that a system may encounter when operating on in-the-wide videos. Given a 3D pose lifter trained on the 2D keypoints estimated from uncorrupted frames, if the RGB input is perturbed at test time, the 2D estimated pose by g_ϕ would be erroneous. This error would *propagate to the lifter module* f_θ and leads to poor generalization as expected. In some applications where privacy-preserving is crucial, the collected input could be just sequences of 2D detected keypoints from some g_ϕ (i.e., improving the robustness of g_ϕ is impossible). Our goal is to train a 2D-to-3D pose lifter, i.e. f_θ, given a sequence of 2D detected keypoints ($\mathbf{x}$) by g_ϕ, that is robust to possible sources of error in the 2D pose input ($\mathbf{v}$), such as occlusion or noise. In previous works [32, 65], f_θ is vulnerable to noisy $\mathbf{x}$ since it is trained with ground truth 2D-to-3D pairs.

Scenarios and Proposed Approaches: In this work, we propose approaches for two different scenarios:

- **Scenario 1:** Corruption operators are *unknown*. f_θ is exposed to an uncorrupted dataset at training while the dataset at testing time is corrupted (i.e., covariate-shift [33]). Our data augmentation strategy for increasing the robustness of f_θ is introduced in Sec. 4.
- **Scenario 2:** Corruption operators are *known* during training. The task is to design f_θ that can efficiently learn from noisy data. We propose a strategy in Sec. 5 to replace parts of f_θ with our confidence-aware module.

4. TAGN: Temporal Additive Gaussian Noise

Motivation: Under the first scenario, image-domain augmentation is the most direct approach. Since g_ϕ is fixed, one can randomly transform the RGB inputs $\mathbf{v}$ by a set of data augmentation operators, mimicking the true corruptions at test time. We then use the 2D detected keypoints of these samples to train f_θ. However, operating on RGB frames is computationally expensive (passing all augmented frames through g_ϕ), and requires having prior knowledge of the corruptions at test time (to design a proper data-augmentation strategy). Meanwhile, a preferred way is augmenting the lower dimension 2D-pose-domain data ($\mathbf{x}$) directly. The drawback is that *it is extremely challenging to model the induced noise in the estimated 2D pose caused by the image-domain corruptions*. This is due to the inherent dependence on the frame appearance, the 2D pose detector, and even the type of video corruption. To support this argument, in Figure 3, we depict the distribution of the errors in the detected 2D keypoints (by HRNet [44]), after corrupting the videos with two different operators (specifically, guided-patch erasing - *top* and Gaussian noise - *bottom*; see Sec. 6.1 for further details). Note how this error distribution varies with the joint and the type of video corruption.

Therefore, we aim to search for an easy-to-simulate and universal noise model, transforming the uncorrupted 2D pose directly. We conjecture that training f_θ with a simple noisy 2D pose input (without needing to approximate the true induced noise) enhances its robustness and indirectly models detected 2D pose of randomly corrupted frames.

TAGN: We design TAGN keeping in mind that in real-world settings, for some periods of time, parts of the human body can be occluded, and therefore their corresponding estimated 2D pose is noisy. For example, consider a

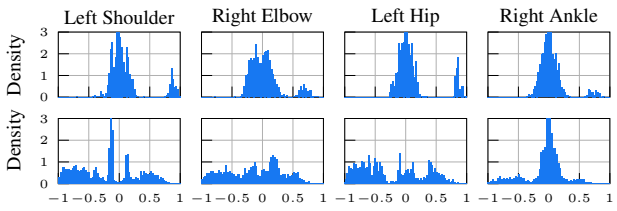

Figure 3. Histogram of the ℓ_2 error of several 2D keypoints detected by HRNet [44], after applying guided patch erasing *(top)* and Gaussian noise *(bottom)* video corruptions defined in Sec. 6.1.

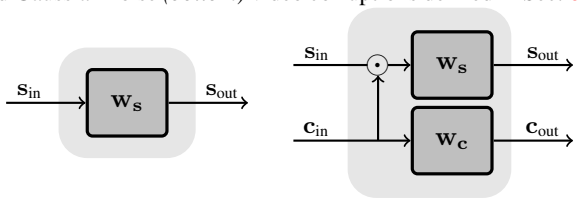

Figure 4. Regular convolution block (left) versus our proposed confidence-aware convolution block (right).

subject walking around an area with several objects on the ground. While passing nearby objects, the subject's feet may become occluded for a short or extended period. As a result, the error in the estimated 2D keypoints by g_ϕ only appears in a portion of 2D joints at some frames. We construct TAGN to closely simulate these situations. Given a 2D pose sequence, TAGN operates by first randomly selecting $k\%$ of the frames. Next, for each candidate frame, $p\%$ of the joints are drawn randomly and contaminated by additive Gaussian noise with $\mathcal{N}(0, \sigma^2)$ distribution. In short, TAGN is a random data augmentation strategy that rather than the RGB input, directly affects the 2D pose input to the lifter. We provide the outline of TAGN in the Supplementary.

Ultimately, one can imagine TAGN to be a simplified error model on the 2D detected keypoint inputs. Despite this, we still found TAGN to be an effective 2D pose data augmentation strategy that improves the robustness to unforeseen video corruptions (as will be shown in Sec. 7.1).

5. CA-Conv: Confidence-aware Convolution

Motivation: Many 2D HPE methods output a confidence score for each detected joint [44, 63]. However, 3D lifters often *ignore* this score and only resort to estimated 2D pose as their input. Here, we aim to incorporate the 2D pose confidence scores in the 3D pose lifting step. To examine this idea, we revise the VideoPose3D [32] architecture by modifying its regular 1D convolution block to be confidence aware, hence the name *confidence-aware convolution* (CA-Conv). The training is performed on noisy keypoints (Scenario 2). This idea can be extended to other architectures.

CA-Conv: Each CA-Conv block contains two 1D convolutional sub-blocks with separate $\mathbf{w_s}$ and $\mathbf{w_c}$ kernels. Furthermore, it consumes two input streams $\mathbf{s}_{\text{in}} \in \mathbb{R}^{L_{\text{in}} \times D_{\text{in}}}$ and $\mathbf{c}_{\text{in}} \in [0,1]^{L_{\text{in}} \times D_{\text{in}}}$, with dimensions $L_{\text{in}} \times D_{\text{in}}$ (sequence length $\times$ embedding dimension, per sample in each mini-batch). While $\mathbf{s}_{\text{in}}$ is the feature map encoding the pose information, $\mathbf{c}_{\text{in}}$ denotes the intermediate features of the confidence scores. The two output streams of each CA-Conv block are obtained following:

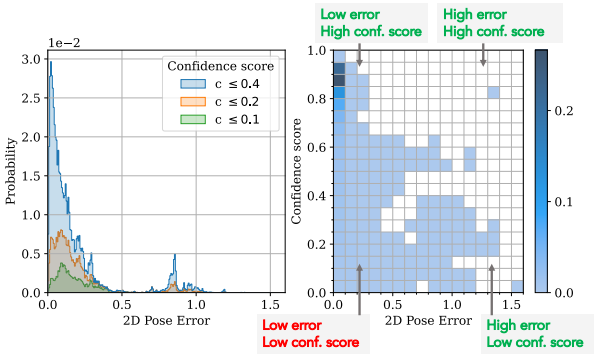

Figure 5. Distribution of the ℓ_2 error of the left shoulder's detected 2D pose, before and after applying guided patch erasing, for different confidence scores c *(left)*. Joint histogram of the error in the detected 2D pose and the confidence score *(right)*. In the right subplot, the color specifies the density. Best viewed in color.

$$\mathbf{s}_{\text{out}}[d] = \text{ReLU}\left(\sum_{i=1}^{D_{\text{in}}} (\mathbf{s}_{\text{in}}[i] \odot (\gamma + \mathbf{c}_{\text{in}}[i])) * \mathbf{w_s}[i,d]\right) \quad (1)$$

$$\mathbf{c}_{\text{out}}[d] = \text{Sigmoid}\left(\sum_{i=1}^{D_{\text{in}}} \mathbf{c}_{\text{in}}[i] * \mathbf{w_c}[i,d]\right) \quad (2)$$

where i and d index the input and output channels, respectively. $\mathbf{s}_{\text{out}}, \mathbf{c}_{\text{out}} \in \mathbb{R}^{L_{\text{out}} \times D_{\text{out}}}$ are output tensors of each CA-Conv block, which are fed to the subsequent blocks. Also, $\odot$ is the element-wise product, while $*$ is the convolution operation. Note that, in (2), we are channel-wise weighting the pose feature maps $\mathbf{s}_{\text{in}}$ by the confidence score features $\mathbf{c}_{\text{in}}$. Also, to avoid the effect of 2D pose features with low confidence scores being completely diminished, we consider a positive leakiness term γ, added to the confidence feature map $\mathbf{c}_{\text{in}}$. Figure 4 compares the CA-Conv versus a regular convolution block.

As there is no explicit supervision on the confidence score when training 2D HPE models, they might be erroneous and thus need to be dealt with cautiously. For instance, we noticed that in HRNet [44], while some 2D keypoints are detected accurately, they might be associated with a relatively low confidence score. To exemplify this, in Figure 5, we visualize the histogram of the ℓ_2-error between the left-shoulder 2D keypoint output by HRNet, with and without video corruption. We treat the 2D pose with no corruption to be an accurate estimation and compare it with the inferred 2D keypoint in the presence of video corruption.

We observe that the keypoints with high confidence scores (≥ 0.8) in Figure 5-right, carry useful information with the error distribution concentrated around zero. This means that most of the keypoints that are accurately detected correspond to a high confidence score. On the other

hand, the confidence scores might not perfectly correlate with the error in 2D keypoint detection. For example, in Figure 5-right, a large portion of 2D keypoint estimations, regardless of their accuracy (i.e. with high and low ℓ_2 errors) have a low confidence score. Furthermore, in Figure 5-left, many less confident keypoint estimations seem to have low detection errors (see the green histogram, concentrated in the low-error region). This implies that ignoring some inferred 2D keypoints simply based on their confidence score can potentially harm the 3D lifter's performance and justifies our leakiness component γ.

6. Experimental Setup

6.1. Corrupted 3D HPE Datasets

Standard 3D HPE Benchmarks: Our evaluation benchmarks are based on the following standard 3D HPE datasets:

- *Human 3.6M (H36M)* [13]: This dataset contains 3.6 million video frames capturing 11 subjects performing 15 actions. Following [32], our training and test sets include (S1, S5, S6, S7, S8) and (S9, S11) subjects, respectively.
- *HumanEva-I* [42]: Compared to H36M, this dataset is much smaller, with only 3 subjects captured. We followed the same evaluation protocol as in [26, 32], where each video is partitioned for training and evaluation purposes. Three actions (Walk, Jog, Box) are used for evaluation.

Image-domain Corruption Operators: To propose new benchmarks for evaluating the robustness of video-based 2D-to-3D pose lifters, we introduce several *image-domain operators* that simulate the common artifacts appearing in videos captured in-the-wild. Following [10, 27, 38], we develop 6 video corruption operators, including guided-patch erasing, temporal patch erasing, pixel-level (Gaussian and impulse) noise, cropping, and synthesized motion blur. Figure 6 illustrates the effect of these operators on example frames from the Human 3.6M [13].

In *guided-patch erasing*, instead of regular random-patch occlusion [39] or directly masking the joint positions [2], we propose a guided erasing strategy to effectively mask out fixed-size square patches in a video. Based on the datasets' ground-truth 2D keypoints, we first uniformly select a subset of joints and track their trajectory throughout a sequence of frames. Next, we choose the masking patch positions such that they have the most overlap with the keypoints' footprint. We also consider an extension of this operator called *temporal patch erasing*, where a video is partitioned into N non-overlapping sequences (N is randomly selected from 1 to 10) and the guided-patch erasing is applied on each shorter sequence independently.

To simulate noise introduced due to low-light conditions or video compression [10], we corrupt the video frames with Gaussian and impulse noises. We also consider *motion blur* operator to account for the possible blurring artifacts due to the motion of the subject or the camera. Finally,

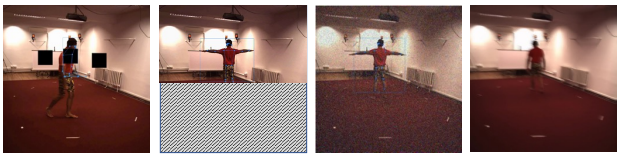

Figure 6. Examples from our Human3.6M-C dataset. More examples will be provided in the Supplementary.

to simulate partial visibility of subjects in a video, we adopt a *cropping* operator that horizontally crops the frames. Implementation details are provided in the Supplementary.

Corrupted Variations of H36M & HumanEva-I: Applying our corruption operators to standard 3D HPE datasets allows us to generate variations of the original datasets (each with the same number of frames). We call this dataset *H36M-Corrupted* (H36M-C), similar to the naming in [10]. The *test* set of H36M-C is generated by performing all video corruptions to the test set of H36M. For the *train* set of H36M-C, each corruption operator is applied to 10% randomly selected videos from the H36M's training set (without replacement). We keep the remaining unselected videos without corruption in the H36M-C dataset. Therefore, the size of the train sets of H36M and H36M-C datasets is the same. Similarly, we create the HumanEva-I-C dataset by applying all corruption operators, except the cropping (for reasons we will discuss later), to the whole HumanEva-I.

6.2. Evaluation Metrics

MPJPE & P-MPJPE: The mean per-joint position error (MPJPE) in mm, computes the Euclidean distance between the ground-truth (GT) and predicted joint positions. P-MPJPE (Procrustes-MJPE), on the other hand, computes MPJPE after the GT and the predicted 3D pose are scale, rotation and translation aligned.

MPJPE$_{\leq\tau}$: In our corrupted datasets, the corruptions affect each joint differently. To maintain fairness and consistency, MPJPE should only be computed on 3D joints with accurate 2D keypoints as input. For example, if the subject's legs were masked out, the estimation of the occluded joints is ambiguous and should be excluded from the evaluation. For each frame, we include the j-th joint in our evaluation, only if the error on its corresponding 2D keypoints before ($\mathbf{x}$) and after ($\widetilde{\mathbf{x}}$) corruption is less than the threshold τ, i.e. $\|\mathbf{x}_j - \widetilde{\mathbf{x}}_j\| \leq \tau$. We denote this metric as MPJPE$_{\leq\tau}$. Analogously, we define P-MPJPE$_{\leq\tau}$. In our experiments, we select $\tau = 0.1$, which leads to, on average, 87% of the joints (per frame) being included in the H36M-C test set. We provide the percentage of counted joints in evaluation as a function of τ in the Supplementary.

6.3. Baseline Models and Benchmark Performance

Our default 2D keypoint detector is HRNet [44] trained on COCO [20]. We found HRNet performing better than com-

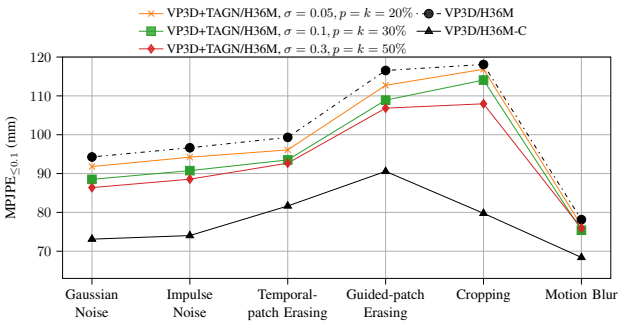

Figure 7. Effect of TAGN with different severity levels. We compare against the lower- and upper-bound performances plotted with ▲ and ● markers, respectively. TAGN outperforms the upper bound. All models are tested on H36M-C.

monly adopted 2D pose estimators (e.g., CPN [5] or Detectron [57]) in 3D pose lifting literature. For 3D lifters, we experiment with several architectures: VP3D [32], PoseFormer [64], SRNet [44], Attention3DHP [21] and Pose3DRIE [40]. We use the default hyper-parameters and training settings [20, 21, 32, 40, 44, 57, 64]. Models are evaluated on the test sets of H36M-C and HumanEva-I-C.

For each lifter, we train two models (from scratch) on (1) the original (H36M, HumanEva-I) and (2) the corrupted (H36M-C, HumanEva-I-C) datasets, serving as baselines for the *two scenarios* (Sec. 3), respectively. In the first scenario, a model is trained on the original dataset with no video corruptions, it generalizes poorly on the corrupted videos. Therefore, we treat its performance as an *upper bound* on MPJPE. On the other hand, the dataset distributions at the training and test time are well-matched for the second scenario. Therefore, its performance is deemed a *lower bound* on MPJPE. For the sake of brevity, we refer to these *benchmark performances* as upper- and lower-bound throughout the rest of this paper. Our approaches are compared against these bounds. A similar extension can be made for lower- and upper-bounds of other metrics such as P-MPJPE. To improve the robustness of the models trained on the original dataset, we adopt TAGN as a 2D pose augmentation. We assess these models against the *upper bound*. We also test the effectiveness of CA-Conv for models trained on corrupted videos and compare them with the *lower bound*. The implementation details of the methods will be elaborated upon in detail in the Supplementary.

7. Results

7.1. Learning with Jittered 2D Pose

TAGN with VP3D: We use TAGN to augment the 2D pose training set of H36M with various (p, k, σ) parameters and use this dataset to train a VP3D model (with a temporal receptive field of 27 frames). Note that, TAGN is only applied once at the beginning of the training process. We repeat each TAGN experiment five times with different noise realizations on the 2D pose input and report the average $\text{MPJPE}_{\leq 0.1}$ evaluated on H36M-C in Table 1 (rows 1-6).

We found a small standard deviation on the performance metrics across all runs in most cases (details in the Supplementary). Despite its simplicity, TAGN empirically improves the lifter's robustness across multiple video corruptions. Noteworthy, the lifters trained with noisy 2D poses are universal and lack knowledge of the exact input RGB video corruption operator applied at test time. The gap between the performance of TAGN under different settings versus the lower- and upper-bounds is visualized in Figure 7. We also consider a baseline in which the 2D keypoints of H36M-C are denoised by a median filter with kernel size 5 (at both train and test times).

Models with TAGN successfully outperform the lower-bound and the simple denoising approach. Besides, the higher level of noise added to the 2D pose during training leads to improved robustness of VP3D models. Noteworthy, TAGN does not introduce surplus complexity during training. Unlike video-level corruptions, we can augment any dataset with additive jitter instantaneously, evade heavy computations, or running inference on 2D pose detectors.

TAGN with Other Pose Lifters: Besides VP3D, in Table 1, we demonstrate the effect of TAGN on other state-of-the-art 2D-to-3D pose lifting solutions. Given the diversity of the studied architectures, we have the same observation regarding the enhanced robustness after using TAGN.

TAGN with VP3D on HumanEva-I: We also evaluate VP3D with TAGN on HumanEva-I [42] in Table 2. Due to the length of videos in this dataset, the distortion ratios (i.e. p and k) of TAGN are reduced to 10%. Furthermore, we found that cropping severely affects the quality of the estimated 2D pose. We attribute this to the fact that in HumanEva-I, the videos have lower resolution and the subjects remain mainly close to the center of the frame. Therefore, horizontal cropping completely removes visual cues of the lower body joints throughout almost all frames, leading to significant errors in 2D pose estimation. Thus, we exclude this video corruption in our analysis in Table 2.

TAGN outperforms the upper-bound on three corruptions (motion blur, Gaussian, and impulse noise). However, the performance on temporal- and guided- patch erasing does not improve for the aforementioned reasons (e.g. lower mobility of the subjects). In our analysis on HumanEva-I, we excluded lifters that require a large training set (e.g., PoseFormer, Attention3DHP) or are unadaptable to a different human skeleton layout (e.g., SRNet).

Qualitative Comparison: In Figure 8, we visualize the effect of TAGN 2D pose augmentation on the quality of the estimated 3D pose. When the input 2D pose is precise or highly inaccurate (column 1), VP3D and VP3D+TAGN perform similarly (columns 2-3). However, we observe the advantage of TAGN, when the errors in 2D input pose *happen*

Table 1. MPJPE$_{\leq 0.1}$ (performance gain compared to the upper bound, lower is better) evaluated on H36M-C. For all tables in the rest of this paper, the mean value across 5 runs is reported for models with TAGN, our methods are highlighted in gray.

Model	Gaussian Noise	Impulse Noise	Temporal-patch Erasing	Guided-patch Erasing	Cropping	Motion Blur	Average
VP3D[32]/H36M *(upper bound)*	94.27	96.64	99.32	116.54	118.08	78.14	100.50
VP3D[32]/H36M + Median Filter	89.55 (↓ 4.72)	92.71 (↓ 3.93)	97.87 (↓ 1.45)	111.86 (↓ 4.86)	118.82 (↑ 0.78)	75.22 (↓ 2.92)	97.65 (↓ 2.85)
VP3D[32]/H36M+TAGN ($\sigma = 0.05; p = k = 20\%$)	91.73 (↓ 2.54)	94.12 (↓ 2.52)	96.08 (↓ 3.24)	112.72 (↓ 3.82)	117.23 (↓ 0.85)	76.23 (↓ 1.91)	98.07 (↓ 2.43)
VP3D[32]/H36M+TAGN ($\sigma = 0.1; p = k = 30\%$)	88.44 (↓ 5.83)	90.69 (↓ 5.95)	93.56 (↓ 5.76)	108.95 (↓ 7.59)	114.3 (↓ 3.78)	75.38 (↓ 2.76)	95.22 (↓ 5.28)
VP3D[32]/H36M+TAGN ($\sigma = 0.3; p = k = 50\%$)	86.3 (↓ 7.97)	88.46 (↓ 8.18)	92.57 (↓ 6.75)	106.74 (↓ 9.8)	107.71 (↓ 10.37)	75.89 (↓ 2.25)	92.94 (↓ 7.56)
VP3D[32]/H36M-C *(lower bound)*	73.11	74.03	81.65	90.56	79.76	68.40	77.92
PoseFormer[64]/H36M *(upper bound)*	103.16	105.09	115.91	132.37	151.62	93.13	116.88
PoseFormer[64]/H36M+TAGN ($\sigma = 0.3; p = k = 50\%$)	102.24 (↓ 0.92)	104.78 (↓ 0.31)	113.8 (↓ 2.11)	129.33 (↓ 3.04)	148.39 (↓ 3.23)	88.7 (↓ 4.43)	114.54 (↓ 2.34)
PoseFormer[64]/H36M-C *(lower bound)*	84.18	85.08	94.61	106.38	96.28	80.37	91.15
SRNet[60]/H36M *(upper bound)*	96.06	98.45	102.36	120.29	126.08	81.27	104.08
SRNet[60]/H36M+TAGN ($\sigma = 0.3; p = k = 50\%$))	91.68 (↓ 4.38)	94.04 (↓ 4.41)	96.16 (↓ 6.2)	111.94 (↓ 8.35)	117.28 (↓ 8.8)	78.7 (↓ 2.57)	98.3 (↓ 5.78)
SRNet[60]/H36M-C (lower bound)	78.17	79.56	84.60	93.94	82.57	71.98	81.8
Attention3DHP[21]/H36M *(upper bound)*	95.13	97.80	100.32	118.48	121.02	77.48	101.7
Attention3DHP[21]/H36M+TAGN ($\sigma = 0.3; p = k = 50\%$)	92.05 (↓ 3.08)	93.94 (↓ 3.86)	100.48 (↑ 0.16)	115.08 (↓ 3.4)	112.8 (↓ 8.22)	84.35 (↑ 6.87)	99.78 (↓ 1.92)
Attention3DHP[21]/H36M-C *(lower bound)*	73.67	74.82	91.96	91.21	79.2	68.24	78.18
Pose3D-RIE [40]/H36M *(upper bound)*	104.32	105.91	108.72	120.05	117.83	94.27	108.51
Pose3D-RIE[40]/H36M+TAGN ($\sigma = 0.3; p = k = 50\%$)	92.14 (↓ 12.18)	101.34 (↓ 4.57)	105.45 (↓ 3.27)	115.31 (↓ 4.74)	104.42 (↓ 13.41)	92.4 (↓ 1.87)	101.84 (↓ 6.67)
Pose3D-RIE [40]/H36M-C *(lower bound)*	86.41	87.56	93.47	97.12	94.52	83.48	90.42

Table 2. MPJPE$_{\leq 0.1}$ and performance gain compared to the upper bound (tested on HumanEva-I-C).

Model	Gaussian Noise	Impulse Noise	Temporal-patch Erasing	Guided-patch Erasing	Motion Blur	Average
VP3D[32]/HumanEva-I *(upper bound)*	66.67	62.05	33.64	33.32	55.82	50.30
VP3D[32]/HumanEva-I+TAGN($\sigma = 0.3; p = k = 10\%$)	59.21 (↓ 7.36)	55.22 (↓ 6.83)	33.74 (↑ 0.10)	33.39 (↑ 0.07)	48.87 (↓ 6.95)	46.11 (↓ 4.19)
VP3D[32]/HumanEva-I-C *(lower bound)*	35.19	33.92	32.33	31.72	33.41	33.31

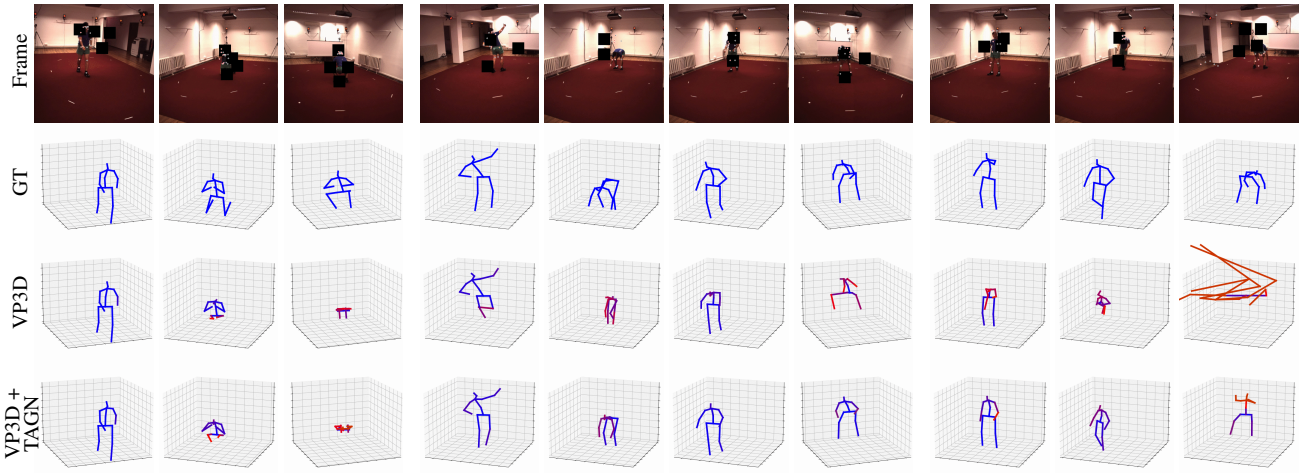

Figure 8. Qualitative comparison of VP3D+TAGN ($\sigma = 0.3, p = k = 50\%$) versus VP3D, on H36M-C dataset (with guided-patch erasing corruption). The bone color leading to a joint turns red when its MPJPE increases.

only on a few keypoints or temporarily (columns 4-7). Remarkably, while VP3D fails in some challenging situations, our VP3D+TAGN still preserves its robustness and outputs a reasonable 3D pose (columns 8-10).

7.2. Learning with Confidence Scores

In Table 3, we showcase the results of VP3D with all regular 1D convolutions replaced by our CA-Conv blocks (VP3D+CA-Conv). For this analysis, we trained on H36M-C. To the best of our knowledge, there is no similar 2D-to-3D human pose lifting approach that directly consumes the output confidence scores from a 2D pose detector. Thus, we assess the performance of VP3D+CA-Conv against two baselines. Notably, we show that a simple extension of VP3D - concatenating the confidence score to the 2D pose input (VP3D+Conf.-Concat) similar to [58] or denoising the 2D keypoints sequence with median filter (with kernel size 5) before passing to the lifter are not effective. We hypothesize that in VP3D+Conf.-Concat, it is difficult to effectively entangle the position and confidence scores, with inherently different scales and meanings, using a single convolution.

7.3. Ablation Study

In Table 4, we evaluate the effect of temporal receptive fields $\{1, 9, 27, 81\}$ on VP3D performance, trained with either TAGN ($\sigma = 0.3$, $p = k = 50\%$) or CA-Conv ($\gamma = 1$),

Table 3. Effect of CA-Conv blocks on MPJPE$_{\leq 0.1}$. All models are trained and tested on H36M-C dataset.

Model	Gaussian Noise	Impulse Noise	Temporal-patch Erasing	Guided-patch Erasing	Cropping	Motion Blur	Average
VP3D[32]/H36M-C *(lower bound)*	73.11	74.03	81.65	90.56	79.76	68.40	77.92
VP3D[32]/H36M-C + *Median Filter*	75.06 (↑ 1.95)	76.14 (↑ 2.11)	82.42 (↑ 0.77)	89.11 (↓ 1.45)	82.04 (↑ 2.28)	70.12 (↑ 1.72)	79.14 (↑ 1.22)
VP3D[32]/H36M-C + *Conf. Concat*	74.42 (↑ 1.31)	75.18 (↑ 1.15)	80.81 (↓ 0.84)	89.52 (↓ 1.04)	81.19 (↑ 1.43)	68.84 (↑ 0.44)	78.33 (↑ 0.41)
VP3D[32]+CA-Conv ($\gamma = 0$)/H36M-C	73.11 (↓ 0.00)	73.96 (↓ 0.07)	79.77 (↓ 1.88)	88.00 (↓ 2.56)	78.83 (↓ 0.93)	67.85 (↓ 0.55)	76.92 (↓ 1.00)
VP3D[32]+CA-Conv ($\gamma = 1$)/H36M-C	72.31 (↓ 0.80)	73.28 (↓ 0.75)	79.45 (↓ 2.20)	87.72 (↓ 2.84)	76.84 (↓ 2.92)	67.28 (↓ 1.24)	73.64 (↓ 1.79)

Table 4. Effect of VP3D's [32] receptive field on MPJPE$_{\leq 0.1}$ and the performance gain compared to lower- or upper-bounds.

Receptive Field	Model	Average MPJPE$_{\leq 0.1}$
1	VP3D/H36M *(upper bound)*	96.87
	VP3D/H36M+TAGN	91.39 (↓ 5.48)
	VP3D/H36M-C *(lower bound)*	83.09
	VP3D/H36M-C + CA-Conv	80.25 (↓ 2.84)
9	VP3D/H36M *(upper bound)*	96.26
	VP3D/H36M+TAGN	92.79 (↓ 3.47)
	VP3D/H36M-C *(upper bound)*	79.37
	VP3D/H36M-C + CA-Conv	76.94 (↓ 2.43)
27	VP3D/H36M *(upper bound)*	94.19
	VP3D/H36M+TAGN	88.54 (↓ 5.65)
	VP3D/H36M-C (lower bound)	75.46
	VP3D/H36M-C + CA-Conv	73.64 (↓ 1.82)
81	VP3D/H36M *(upper bound)*	93.35
	VP3D/H36M+TAGN	87.69 (↓ 5.66)
	VP3D/H36M-C *(lower bound)*	73.61
	VP3D/H36M-C + CA-Conv	71.84 (↓ 1.77)

Table 5. MPJPE$_{\leq 0.1}$ of VP3D [32] models trained/tested on 2D keypoints detected by HRNet [44] and Lite-HRNet [59].

Training 2D keypoints	Model	Testing 2D keypoints	
		HRNet	Lite-HRNet
HRNet	VP3D/H36M *(upper bound)*	100.50	118.45
	VP3D/H36M+TAGN	92.94	112.26
	VP3D/H36M-C *(lower bound)*	77.92	98.03
	VP3D/H36M-C + CA-Conv	76.13	99.31
Lite-HRNet	VP3D/H36M *(upper bound)*	101.66	115.94
	VP3D/H36M+TAGN	98.07	110.70
	VP3D/H36M-C *(lower bound)*	84.08	94.16
	VP3D/H36M-C + CA-Conv	81.59	89.01

against its baseline. We observe a consistent performance boost for different receptive field sizes when TAGN or CA-Conv are adopted. This demonstrates our approaches are suitable for various choices of this parameter.

We examine the effect of the 2D detector in Table 5. In addition to HRNet, we also employ Lite-HRNet [59] - a lightweight version of HRNet - which estimates 2D pose in real-time. Due to the simplicity of Lite-HRNet, the detected 2D keypoints are less accurate and generally result in a higher 3D lifting error. Regardless, the models with TAGN and CA-Conv still outperform the upper- and lower-bounds. Moreover, we also do cross-detector evaluations, where the 2D keypoints detector at training differs from the one at test time. When VP3D/H36M-C+CA-Conv is trained with HRNet 2D pose, it is exposed to more accurate 2D input pose and confidence scores compared to Lite-HRNet (see the Supplementary for a comparison of Lite-HRNet and HRNet confidence score distribution). This slightly affects the performance when tested on Lite-HRNet keypoints but not the other way around. In general, CA-Conv expects similar or higher quality test-time 2D keypoints and confidence scores to achieve superior robustness.

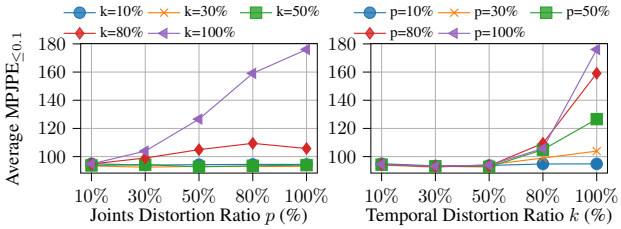

Figure 9. Effect of TAGN's joint and temporal distortion ratios (p and k) with $\sigma = 0.3$. Here, the lifter architecture is VP3D.

We analyze the impact of TAGN's temporal and joint distortion ratios (i.e. k and p) in Figure 9. While the joint distortion ratio (p) does not visibly impact the performance, increasing the temporal distortion ratio (k) gradually degrades the results, especially when more than half of the frames are affected. This highlights the effectiveness of TAGN and the fact that blindly adding Gaussian noise to every joint and frame leads to the worst performance since the model cannot rely on any joint or frame within its receptive field.

8. Conclusion

By rigorously studying the robustness of 2D-to-3D pose lifters under two introduced benchmarks with video corruptions, we showed in spite of accurate 3D pose estimation of state-of-the-art 2D-to-3D pose lifters on standard benchmarks, they are extremely sensitive to the unexpected distortions in their input. This observation prompts the necessity of careful investigation of robustness in this task. We took steps towards improving the robustness of lifter models by introducing (1) TAGN - a simple 2D pose augmentation approach, (2) infusing the knowledge of the 2D detector's confidence score at the 3D lifter side through our confidence-aware convolution block design. Our extensive experimental results revealed the effectiveness of our methods in boosting the robustness of 3D pose lifters.

Acknowledgement

This project has been funded by the Jump ARCHES endowment through the Health Care Engineering Systems Center, JSPS/MEXT KAKENHI (24K20830), and ROIS NII Open Collaborative Research (2023-23FC01, 2024-24S1201).

References

[1] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Tom Cashman. Flag: Flow-based 3D avatar generation from sparse observations. In *2022 Computer Vision and Pattern Recognition*, June 2022. 1

[2] Amin Ansarian and Maria A. Amer. Realistic augmentation for effective 2D human pose estimation under occlusion. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 919–923, 2021. 3, 5

[3] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. *Advances in Neural Information Processing Systems*, 33:20496–20507, 2020. 2

[4] Ching-Hang Chen and Deva Ramanan. 3D human pose estimation = 2D pose estimation + matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6

[6] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3D human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10631–10638, 2020. 2

[7] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby Tan. Occlusion-aware networks for 3D human pose estimation in video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 723–732, 2019. 2

[8] Wojciech M. Czarnecki and Igor T. Podolak. Machine learning with known input data uncertainty measure. In Khalid Saeed, Rituparna Chaki, Agostino Cortesi, and Sławomir Wierzchoń, editors, *Computer Information Systems and Industrial Management*, pages 379–388, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 2

[9] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8575–8584, 2021. 2

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 2, 5

[11] Trung-Hieu Hoang, Mona Zehni, Huaijin Xu, George Heintz, Christopher Zallek, and Minh N. Do. Towards a comprehensive solution for a vision-based digitized neurological examination. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4020–4031, 2022. 1

[12] Wenbo Hu, Changgong Zhang, Fangneng Zhan, Lei Zhang, and Tien-Tsin Wong. Conditional directed graph convolution for 3D human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 602–611, 2021. 2

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 1, 3, 5

[14] Tao Jiang, Necati Cihan Camgoz, and Richard Bowden. Skeletor: Skeletal transformers for robust body-pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3402, 2021. 1, 2

[15] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *Int. J. Comput. Vis.*, 129(2):462–483, 2021. 2

[16] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2

[17] Sijin Li and Antoni B Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014. 2

[18] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3D human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[19] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3D human pose estimation. *IEEE Transactions on Multimedia*, 2022. 2

[20] 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. 5, 6

[21] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020. 1, 2, 6, 7

[22] Diogo C. Luvizon, David Picard, and Hedi Tabia. 2D/3D pose estimation and action recognition using multitask deep learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018. 1

[23] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2

[24] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 1

[25] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild

using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2

[26] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3D human pose estimation with a single RGB camera. In *ACM Transactions on Graphics*, volume 36, July 2017. 5

[27] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 2, 5

[28] Francesc Moreno-Noguer. 3D human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2823–2832, 2017. 2

[29] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3D human pose estimation using convolutional neural networks with 2D pose information. In *European Conference on Computer Vision*, pages 156–169. Springer, 2016. 2

[30] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3D human pose estimation using convolutional neural networks with 2D pose information. In *European Conference on Computer Vision*, pages 156–169. Springer, 2016.

[31] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2

[32] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[33] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009. 3

[34] Umer Rafi, Juergen Gall, and Bastian Leibe. A semantic occlusion model for human pose estimation from a single depth image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 67–74, 2015. 2

[35] Russell Reed, Se young Oh, and Robert J. Marks. Regularization using jittered training data. In *Proceedings of IJCNN International Joint Conference on Neural Networks*, volume 3, pages 147–152 vol.3, 1992. 2

[36] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3D human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[37] Chris Rockwell and David F Fouhey. Full-body awareness from partial observations. In *European Conference on Computer Vision*, pages 522–539. Springer, 2020. 2

[38] István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. How robust is 3D human pose estimation to occlusion? IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18) - Workshop on Robotic Co-workers 4.0: Human Safety and Comfort in Human-Robot Interactive Social Environments, October 2018. 1, 2, 3, 5

[39] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. Synthetic occlusion augmentation with volumetric heatmaps for the 2018 ECCV posetrack challenge on 3D human pose estimation. *arXiv preprint arXiv:1809.04987*, 2018. 3, 5

[40] Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3D human pose estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3446–3454, 2021. 2, 6, 7

[41] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3D human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2325–2334, 2019. 2

[42] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, Mar. 2010. 1, 5, 6

[43] Jan Stenum, Kendra M Cherry-Allen, Connor O Pyles, Rachel D Reetzke, Michael F Vignos, and Ryan T Roemmich. Applications of Pose Estimation in Human Health and Performance across the Lifespan. *Sensors*, 21(21), 2021. 1

[44] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3, 4, 5, 6, 8

[45] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 529–545, 2018. 2

[46] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021. 2

[47] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3D people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13243–13252, 2022. 2

[48] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and P. Fua. Learning to fuse 2D and 3D image cues for monocular body pose estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3961–3970, 2017. 2

[49] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3D body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–1000, 2016.

[50] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[51] Bastian Wandt, James J Little, and Helge Rhodin. Elepose: Unsupervised 3D human pose estimation by predicting camera elevation and learning normalizing flows on 2D poses.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6635–6645, 2022. 2

[52] Danding Wang, Wencan Zhang, and Brian Lim. Show or suppress? managing input uncertainty in machine learning model explanations. *Artificial Intelligence*, 294:103456, 01 2021. 2

[53] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11850–11859, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. 1, 2, 3

[54] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, 2021. 1

[55] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3D pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 2

[56] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[57] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[58] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 7

[59] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-HRnet: A lightweight high-resolution network. In *CVPR*, 2021. 8

[60] Ailing Zeng, Xiao Sun, Fuyang Huang, Minhao Liu, Qiang Xu, and Stephen Ching-Feng Lin. SRNet: Improving generalization in 3D human pose estimation with a split-and-recombine approach. In *ECCV*, 2020. 1, 2, 7

[61] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3D pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11436–11445, 2021. 2

[62] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7374–7383, 2020. 2

[63] Tomas Simon Zhe Cao, Gines Hidalgo and Yaser Sheikh Shih-En Wei. Openpose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 3, 4

[64] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D human pose estimation with spatial and temporal transformers. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 6, 7

[65] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11656–11665, 2021. 2, 3

[66] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 2

[67] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 2