# Towards Robust 3D Pose Transfer with Adversarial Learning

Haoyu Chen[1] Hao Tang[2] Ehsan Adeli[3] Guoying Zhao[1,3]*
[1]CMVS, Finland [2]CMU, USA [3]Stanford University, USA
*Corresponding Author
{chen.haoyu, guoying.zhao}@oulu.fi, bjdxtanghao@gmail.com, eadeli@stanford.edu

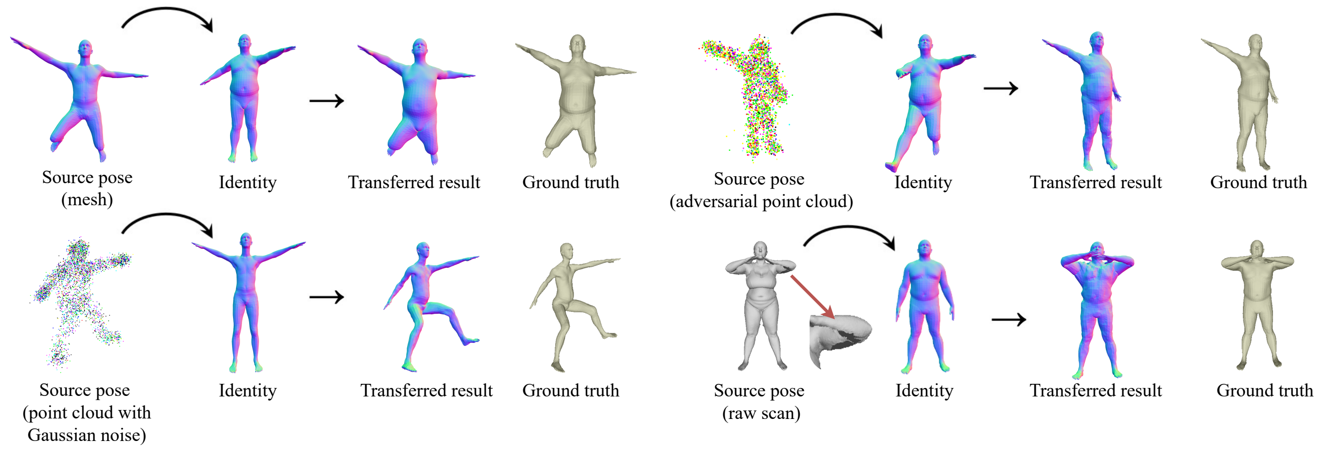

Figure 1. Examples of our 3D pose transfer results on various pose sources, show strong robustness and generalizability. The pose source includes clean mesh (top left) and point clouds with Gaussian noise (bottom left) from SMPL-NPT dataset [41], the adversarial sample of point cloud generated by our method (top right), and raw scan (bottom right) from DFAUST dataset [5]. Identity meshes are from the SMPL-NPT dataset [41] and the FAUST [3] (bottom right) dataset. Our method can achieve promising pose transfer performance even on the extremely challenging *incomplete* raw scan (bottom right). See more results and details in the Supplementary Materials.

## Abstract

*3D pose transfer that aims to transfer the desired pose to a target mesh is one of the most challenging 3D generation tasks. Previous attempts rely on well-defined parametric human models or skeletal joints as driving pose sources. However, to obtain those clean pose sources, cumbersome but necessary pre-processing pipelines are inevitable, hindering implementations of the real-time applications. This work is driven by the intuition that the robustness of the model can be enhanced by introducing adversarial samples into the training, leading to a more invulnerable model to the noisy inputs, which even can be further extended to directly handling the real-world data like raw point clouds/scans without intermediate processing. Furthermore, we propose a novel 3D pose Masked Autoencoder (3D-PoseMAE), a customized MAE that effectively learns 3D extrinsic presentations (i.e., pose). 3D-PoseMAE facilitates learning from the aspect of extrinsic attributes by simultaneously generating adversarial samples that perturb the model and learning the arbitrary raw noisy poses via a multi-scale masking strategy. Both qualitative and quantitative studies show that the transferred meshes given by our network result in much better quality. Besides, we demonstrate the strong generalizability of our method on various poses, different domains, and even raw scans. Experimental results also show meaningful insights that the intermediate adversarial samples generated in the training can successfully attack the existing pose transfer models.*

## 1. Introduction

As a promising and challenging task, 3D pose transfer has been consistently drawing research attention from the computer vision community [10, 27, 37, 41]. The task aims at transferring a source pose to a target identity mesh and keeping the intrinsic attributes (i.e., shape) of the identity mesh. Aside from pure research interests, transferring desired poses to target 3D models has various potential applications in the film industry, games, AR/VR, etc [8, 11, 12].

To achieve data-driven learning, existing 3D pose transfer methods rely on different prerequisites to the data sources, which severely limits their further real-world implementations. Firstly, many existing 3D pose transfer methods [30, 45] cannot directly be generalized to unseen

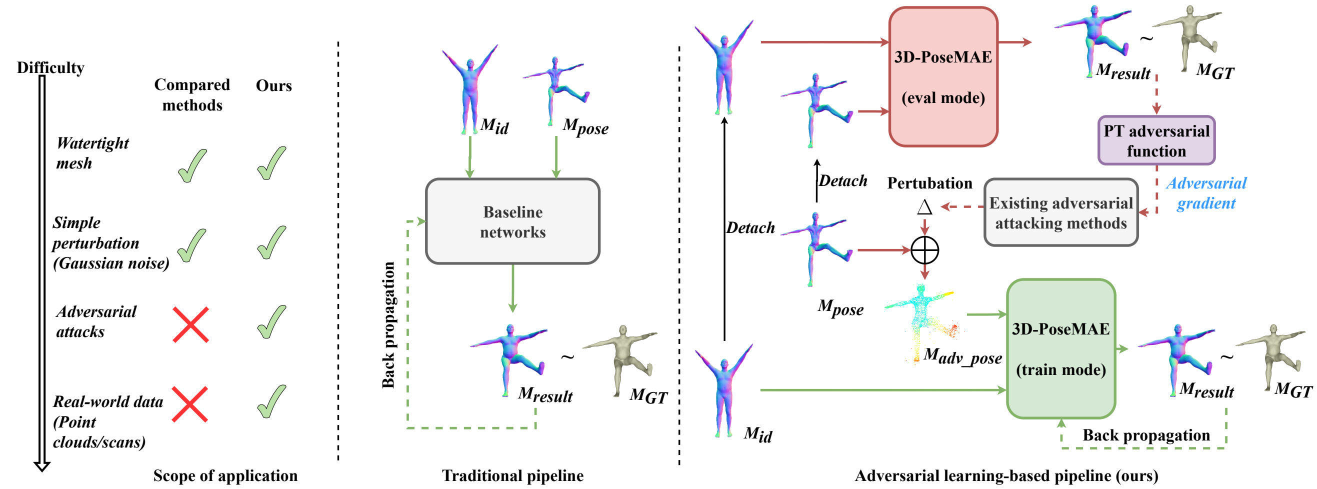

Figure 2. **Left:** The existing methods can deal with simple perturbations such as Gaussian noises but fail to handle harder inputs in real-world cases. **Middle:** The traditional pipeline used in previous methods [10, 27, 37, 41] for 3D pose transfer. The model is trained with clean mesh inputs without considering the robustness to the noisy inputs. We use the symbol $\sim$ to generally refer to the loss term, which differs according to the actual condition. **Right:** Our method. Our method utilizes the strength of adversarial learning to enhance the robustness and generalizability of the model. It consists of an adversarial sample generating flow(top part in red) and a pose transferring flow(bottom part in green). The two flows happen iteratively during the adversarial training and the adversarial samples are calculated on-the-fly. Note that $M_{id}$, $M_{pose}$, $M_{result}$, and $M_{GT}$ stand for the identity, pose, generated meshes, and ground truths, the same as below.

target meshes, and training on the target meshes is inevitable for them to learn the priors of the target shape. Secondly, some studies [1, 19, 27] assume that the paired correspondences between the pose and identity meshes are given, which also involves extra manual efforts to obtain. Lastly, all previous attempts of 3D pose transfer rely on pre-processed and clean source poses to drive the target meshes [41, 51]. However, acquiring clean data is cumbersome and necessary pre-processing pipelines are inevitable. For instance, to register raw human scans to well-defined parametric human models (e.g., SMPL series [4, 29, 34]), it will take roughly 1-2 minutes to process merely a single frame [44], hindering the real-time implementations.

Inspired by the scaling successes of adversarial learning in the CV community [2, 7, 15, 18, 21, 25] for the robustness of the models, we experiment with applying adversarial training to 3D pose transfer tasks. As shown in Fig. 2, Existing methods [10, 41] already show certain robustness to the noisy input (Gaussian noise), however, experimental results show that they are still vulnerable when we directly apply them to real-world point clouds or scan inputs. We suspect there is a domain gap between the synthesized noises and real-world data distribution. Thus, as shown in Fig. 1, we wish to utilize the strength of adversarial learning to enhance the robustness and generalizability of the model with more challenging adversarial attacks, making it go beyond so that conducting pose transfer on unseen domains or even directly from raw scans can be possible.

However, although the idea is intuitive, *it's not feasible to naively extend existing adversarial training algorithms [40, 47, 50] to the 3D pose transfer task*. Primarily, the current methods [46, 50] generate adversarial samples based on discriminative adversarial functions, and to our knowledge, there is no adversarial attack proposed specifically for the 3D generative tasks (i.e., how to quantitatively justify if a sample harms the generated results). Moreover, previous approaches using 3D adversarial samples, such as [46, 50], do so by taking them as pre-computed input data and leaving them untouched for the entire training procedure. This protocol is not practical for our task, as models need to learn the latent pose space via gradients. Thus, we proposed an adversarial learning framework with a new pose transfer (PT) adversarial function and on-the-fly computation of adversarial samples, which enables the successful application of adversarial training for the generative models.

Another novel ingredient in this paper is inspired by the recent powerful capability of masked autoencoding (MAE) architectures [20] in the computer vision community. We adopt the idea of MAE to 3D pose transfer task by implementing a new model, called 3D-PoseMAE, that empathizes the learning of extrinsic presentations (i.e., pose). Specifically, unlike any existing 3D MAE-based models [32, 49, 52] that merely put efforts into depicting spatial local regions to capture the geometric and semantic dependence, 3D-PoseMAE exploits a multi-scale masking strategy to aggregate consistent sampling regions across scales for robust learning of extrinsic attributes. Besides, we observe that local geometric details (wrinkles, small tissues, etc.) from the pose sources are intuitively unessential for pose learning. Thus, we argue that traditional 3D spatial-wise attention/correlation operation [10, 37] might involve a lot of redundant geometric information for pose learning.

Instead, we adopt a progressive channel-wise attention operation to the 3D-PoseMAE so that the attention is operated gradually and fully on the latent pose code, making the presentations more compact and computationally efficient.

The contributions are summarized as follows:

- We work on the robustness problem of 3D pose transfer. To our knowledge, it is the first attempt made to approach 3D pose transfer from the aspect of adversarial learning. As a result, we provide a new research entry that generates adversarial samples to simulate noisy inputs and even raw scans so that conducting ***end-to-end*** pose transfer on raw scans and point clouds is made possible.
- We introduce a novel adversarial learning framework customized for the 3D pose transfer task with a novel PT adversarial function and on-the-fly computation of adversarial samples in backpropagation. It's the first time that on-the-fly computation of adversarial samples appears in a 3D generative deep learning pipeline. *Notably, we show that our proposed adversarial function can be easily plugged into other existing adversarial attack methods*.
- We propose a novel MAE-based architecture for 3D pose transfer with carefully designed components to capture the extrinsic attributions with a multi-scale masking strategy and a progressive channel-wise attention operation. The 3D-PoseMAE shows encouraging performances in both computational efficiency and generative ability.
- Intensive experimental results on various datasets and data sources show that our proposed method achieves promising performances with substantial robustness to noisy inputs and the generalizability to noisy raw scans from the real world. Code will be made available.

## 2. Related Work

**Data-driven 3D Pose Transfer.** Data-driven 3D pose transfer aims to transfer given source poses to target shapes by learning the correspondence between the pose and shape automatically. On the one hand, parametric human model-based methods could bring impressive generated results [30, 45, 56], but their superb performances largely rely on the need for the priors of the target shape, which limits their generalization ability to unseen target meshes. Besides, registering raw human scans to well-defined parametric human models is also cumbersome and time-consuming [44]. On the other hand, some linear blending skinning (LBS) based works [1, 27] can be extended to unseen target meshes, but they assume that annotated landmarks/mesh points or T-posed target meshes are given, which is also a strong condition. Lastly, to our knowledge, all existing methods are within a strong prerequisite that pre-processed and clean source poses are available to drive the target meshes [10, 13, 19, 37, 41]. However, to our knowledge, no effort has been made to study the robustness of the 3D pose transfer to the noisy inputs and even raw scans.

**3D Adversarial Learning.** Adversarial attacks [7, 15] have drawn considerable research attention in the 2D vision models as their severe threat to real-world deployments since it was proposed [18]. When it comes to the 3D field, the attack-generating algorithms could be roughly sorted into Carlini& Wagner (C&W) [6], and Projected Gradient Descent (PGD) [18] groups. C&W-based attacks [40, 47, 50] switch the min-max trade-off problem of adversarial training into jointly minimizing the perturbation magnitude and adversarial loss of attacks. But C&W attacks all suffer from time-consuming issues due to the binary search and optimization iteration. Meanwhile, PGD attacks [16, 28] set the perturbation magnitude as a fixed constraint in the optimization procedure, which can achieve the attack in a much shorter time. However, the attack form is limited to point-shifting, unlike C&W attacks that can perform adding or dropping operations on the point clouds. On the other hand, although some efforts have been made in related tasks such as pose estimation, and motion retargeting [42, 51], there is no strategy proposed specifically for attacking the 3D generative tasks, including pose transfer (i.e., it is unaddressed how to define a successful attack to a generated point cloud/mesh). Besides, research efforts [26, 39, 46] have been made to defend against those attacks on point cloud data for various tasks. However, those approaches conduct the defense training directly on 3D pre-computed adversarial samples. To our knowledge, there is no existing 3D deep learning pipeline that jointly generates adversarial samples and learns the defense of generative models yet.

**Deep Learning Models on Point Cloud.** The pioneering and representative deep learning models on point cloud include PointNet, PointNet++, and Dynamic Graph CNN (DGCNN) [35, 36] as being widely used as benchmark models on various 3D tasks. In the past two years, with the trend of Transformer-based and MAE-based architectures [20] in the computer vision community, many 3D-variant models [32, 48, 49, 52, 54] have been proposed. An analogy to the patches in the ViT [17] and MAE [20] in the 2D field, existing 3D MAE-based architectures [53] represent a point cloud as a set of point tokens/proxies, making it into a set-to-set translation problem. Attention operators will be applied to depict different spatial local regions to better capture the local geometric and semantic dependence for the tasks of 3D classification, semantic segmentation, etc. We argue that the 3D pose transfer task is different, and the learning focus of the pose source should be made on the extrinsic presentations (i.e., pose). Meanwhile, the traditional 3D spatial-wise attention/correlation operation [10, 37] can capture much intrinsic (i.e., shape) information, which is inevitable for tasks like 3D classification, and semantic segmentation. However, this is partially inefficient for the task of 3D pose transfer as the detailed geometric information is redundant for learning the pose correlations.

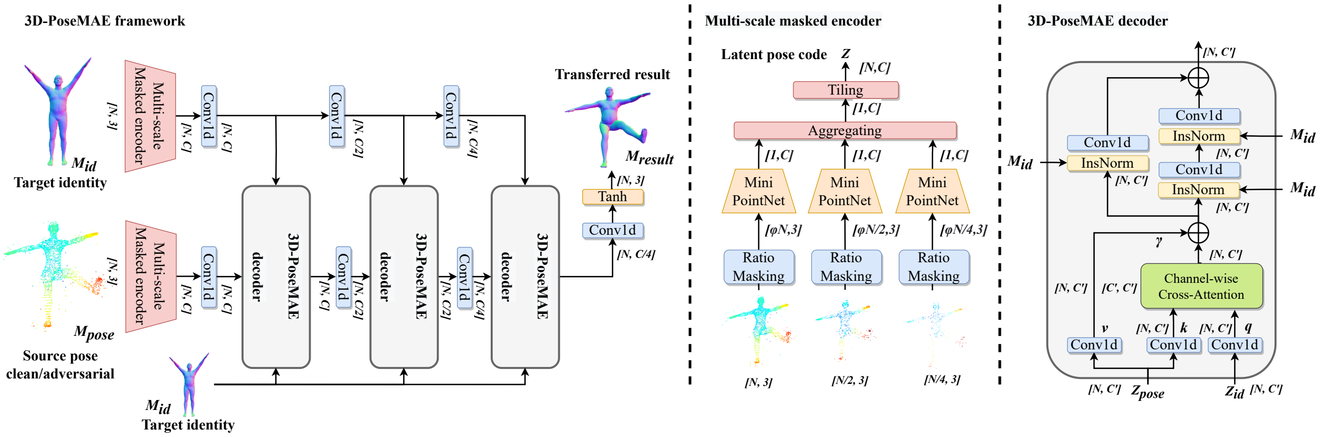

Figure 3. An overlook of our 3D-PoseMAE. The left part is the whole architecture of the 3D-PoseMAE. The middle and right parts illustrate the architectural details of one multi-scale masked encoder and one 3D-PoseMAE decoder, respectively. The 3D-PoseMAE borrows the idea from the work of [20] but is extensively extended to 3D data processing, especially for the 3D pose transfer task. Note that $Z$ stands for the encoded pose feature. $Z_{pose}$ and $Z_{id}$ stand for the specific encoded pose features from pose and identity. Subscripts are the dimensional shape of variables.

# 3. Methodology

We first present a general introduction to the whole pipeline of the proposed method as shown in Fig. 2. Compared to the traditional 3D pose transfer pipeline, we introduce adversarial learning. As shown in Fig. 2, our method consists of an adversarial sample generation procedure (top part in red flow) and a pose transferring (bottom part in green flow) procedure. In the top adversarial sample generation flow, the proposed 3D-PoseMAE model will be set as *eval* mode for obtaining the gradient of the data. By using the gradient of the output of 3D-PoseMAE, we can obtain the perturbation to the meshes, resulting in adversarial samples. In the bottom pose transferring flow, the 3D-PoseMAE model works the same as traditional 3D pose transfer models in a *train* mode, but the pose inputs are replaced with adversarial samples, leading to the adversarial training. Note that the inputs $M_{pose}$ and $M_{id}$ of 3D-PoseMAE need to be detached to ensure the backpropagation of the two flows works without interference.

Next, the overall architecture of the 3D-PoseMAE network together with each component will be introduced, followed by the implementation details of adversarial training.

## 3.1. 3D-PoseMAE

An overview of the 3D-PoseMAE is in Fig. 3. Similar to previous works [10, 37, 41], 3D-PoseMAE takes a source pose (in the form of a mesh/point clouds) and a target mesh as input, and outputs a corresponding pose transferred result. In the following, we introduce each component of the 3D-PoseMAE in an order following the data stream.

**Multi-scale Masked 3D Encoder.** Analogy to the original MAE [20], existing 3D MAE-based architectures [32, 49, 52] invariably chose to split a given point cloud into multiple subsets to conduct the masking. However, this protocol involves feature extraction on each subset for later attention and aggregation, which is computationally demanding and only works on small-scale meshes/point clouds (up to 8,192 points). It cannot be extended to tasks with large-size meshes, such as 3D pose transfer, where the size of the target mesh can be more than 27,000 vertices. Thus, we propose a novel multi-scale masking strategy to boost extrinsic attribute learning. Specifically, as shown in Fig. 3, we regard the input point cloud $M_{pose}$ as the 1-th scale. For the $i$-th scale, $1 \leq i \leq 3$, we randomly downsample the points by $2^{(i-1)}$ times, resulting in a group of the $S$-scale representations of the source pose. Then, we adopt masking at a ratio of $\phi$ to the point cloud at each scale, so that the model will be pushed to learn the same extrinsic attribute shared from those scaled representations via Mini-pointNet [36]. Next, we aggregate the resulting latent pose code from all scales and form the embedding vector $Z$ with dimension $C$. At last, we tail the size of $Z$ from $[1, C]$ to $[N, C]$ for better integration with the identity mesh (N is the vertex number of the identity mesh, which is flexible according to given target meshes) and feed it into the 3D-PoseMAE decoders.

**3D-PoseMAE Decoder with Channel-wise Attention.** After obtaining the latent codes of pose and identity $Z_{pose}$ and $Z_{id}$, we need to adopt the Transformer backbone to integrate two latent codes and conduct the decoding and generating. As there are many existing 3D transformer backbones [10, 32, 54] and the work of GC-Transformer [10] is proposed specifically for 3D pose transfer, we implement our 3D-PoseMAE decoder based on GC-Transformer architecture as shown in the right part of Fig. 3. However, our 3D-PoseMAE decoder has a core difference from any existing 3D Transformer, which is the design of the attention operation. To learn the correlations of given meshes, previous models conduct the attention operation over the spatial channel to perceive the geometric information from the two meshes. Our intuition is that many redundant local geometric representations (wrinkles, small tissues) from the source pose are not essential for pose learning. In other words, we only need the compact pose representations from source poses, it would be encouraging to conduct channel-wise attention where the integration will be fully achieved on the compact pose spaces.

With the above observation, we propose a channel-wise cross-attention module in our 3D-PoseMAE decoder. Firstly, we construct the channel-wise attention map $A$ between $\mathbf{q}$ and $\mathbf{k}$ with the following formula (the subscripts indicate the matrix size):

$$\mathbf{A}_{C' \times C'} = \mathrm{Softmax}(\mathbf{q}^T_{C' \times N} \mathbf{k}_{N \times C'}), \tag{1}$$

where the representations $\mathbf{qk}$ are generated from embedding vectors via different 1D convolution layers (the same as $\mathbf{v}$ below), with the same size as $[N, C']$ ($N$ for vertex number and $C'$ for the channel dimension in the current decoder). Thus, the size of the resulting attention map becomes $[C', C']$. Next, we can obtain the refined latent embedding $Z'_{pose}$ with following

$$Z'_{pose} = \gamma \mathbf{A}_{C' \times C'} \mathbf{v}^T_{C' \times N} + Z_{pose}, \tag{2}$$

where $\gamma$ is a learnable parameter. The remaining design of the 3D-PoseMAE decoder is consistent with the GC-Transformer [10], and the whole network structure is presented in Fig. 3. Please refer to the Appendix for more details on network design and parameter setting.

As we can see, by implementing such a channel-wise attention operation, the network can enjoy two major benefits. Firstly, the size of the intermediate attention map in the network changes from $[N, N]$ to $[C', C']$. Thus, the model size will be substantially reduced. Especially when the processing mesh size is huge, e.g., vertex number $N$ could be up to 27,000 while channel size $C'$ is fixed no more than 1,024 in practice. Secondly, the integration of the pose and target information will be more compact with the cross-attention conducted fully channel-wise, avoiding touching the redundant spatial information brought from the source pose.

**Preserving Fine-grained Geometry.** Note that, to preserve the fine-grained spatial geometric information of the target meshes, we modify the traditional normalization layer into InstanceNorm layer [22, 41]. By the gradual integration in the 3D-PoseMAE, the fine-grained geometry can be refined by the compact pose representations from the channel-wise attention mechanism, see more in the Appendix.

**Optimization**. To train the 3D-PoseMAE for the pose transfer task, we define the full objective function as below:

$$\mathcal{L}_{full} = \mathcal{L}_{rec} + \lambda_{edge} \mathcal{L}_{edge}, \tag{3}$$

where $\mathcal{L}_{rec}$ and $\mathcal{L}_{edge}$ are the two losses used as our full optimization objective, as reconstruction loss and edge loss, the same as following existing baselines [37, 41]. $\lambda_{edge}$ is the corresponding weight of edge loss.

### 3.2. Adversarial Training

An overview of our adversarial learning-based pipeline is on the right side of Fig. 2. Below, we introduce the motivation and problem definition of the adversarial training in the pose transfer task, followed by the implementing details.

**Motivation.** It's proven in various computer vision tasks [2, 7, 21, 43, 57] that introducing adversarial samples to perturb the neural networks during the training can enhance networks' robustness and increase their generalizability and adaptability to other/unseen domains. It's intuitive to think of applying a similar scheme to the 3D pose transfer task so that transferring poses from unseen domains or even directly from raw scans can be made possible. However, this would not work naively by transferring the existing 3D adversarial attack methods to our task due to several issues. Below, we will discuss each issue and illustrate how we approach to the solution.

**Problem Definition.** We define the problem of adversarial training in the 3D pose transfer task as follows. Given a source pose $M_{pose}$ and an identity mesh $M_{id}$, let $\digamma$ be the target pose transfer model (e.g., 3D-PoseMAE) with parameters $\theta$. Then $\digamma(M_{pose}, M_{id}; \theta)$ is the pose transferred mesh generated by the model. For an ideal model we should get: $\digamma(M_{pose}, M_{id}; \theta) = M_{GT}$, with $M_{GT}$ as the ground truth mesh. Then, the problem is converted to an inequality problem: adversarial attack method $f$ needs to generate an adversarial sample $M_{adv\_pose} = f(M_{pose})$ that can satisfy:

$$\digamma(M_{adv\_pose}, M_{id}; \theta) \neq M_{GT}. \tag{4}$$

To solve the inequality problem analogy to Eq. (4), existing 3D adversarial attack methods for 3D object classification tasks [16, 23, 40, 47, 50, 55] convert the inequality into a minimal optimization problem with adversarial loss term $||\mathrm{argmax}_c \digamma(x_{adv}; \theta)_c - t||$ by forcing the prediction to an adversarial sample $x_{adv}$ close to a target class $t$ which is different than the correct one, resulting a successful attack. This is known as targeted attacks. However, this discriminative-based adversarial function cannot directly be applied to generative tasks (e.g., 3D pose transfer) as a continuous latent space is required to present the pose code.

**Untargeted Adversarial Attack for Pose Transfer.** Some non-targeted attack methods [31, 38] convert the above adversarial loss term for targeted attacks into a *reversed* form of $||\mathrm{argmax}_c \digamma(x_{adv}; \theta)_c - t||$ as a minimal optimization problem by pushing away the prediction from the correct class. Inspired by this, we construct *a novel PT adversarial* *function for our 3D generative task in* an intuitive way:

$$f_{adv} = ||\mathcal{F}(M_{pose}, M_{id}; \theta) - M_{GT}||^{-1}. \tag{5}$$

By minimizing the above term, we can push the generated results from the model away from the ground truth mesh, resulting in an attack effect, which is proposed for the first time for the 3D pose transfer.

**The Magnitude of Attacks.** To guarantee the adversarial sample is visually similar to the clean data, adversarial attack methods [16, 23, 40, 47, 50, 55] deploy norms (C&W based attacks) or predefined threshold budget (PGD-based methods) to restrict the perturbations small enough. While in our case, the perturbation ought be strong enough so that the model can handle it without the need to consider the invisibility issue of the perturbation since the goal is not to obtain effective attacks but a robust model. While the magnitude of the adversarial attack is too large, it can be easily defended by simple pre-processing such as statistical outlier removal (SOR) [46]. Thus, to make it closer to real-world applications, we add SOR as pre-processing to all the adversarial samples to filter out those easy samples.

**Adversarial Training for Robustness.** Lastly, after confirming the adversarial loss function as Eq. (5), we can merge it into the existing adversarial attack methods to generate adversarial pose meshes for 3D pose transfer:

$$\begin{aligned} M_{adv\_pose} = M_{clean\_pose} + \\ \phi[f_{adv}(\mathcal{F}(M_{pose}, M_{id}; \theta), M_{GT})], \end{aligned} \tag{6}$$

where $\phi$ can be any existing adversarial attack methods, such as PGD or FGM [28, 47] to generate the perturbations. As mentioned, C&W-based attacks can provide better adversarial samples with less visibility. However, they all suffer from time-consuming issues due to the binary search and heavy optimization iteration. According to our preliminary implementation, merging C&W-based attacks [47] (the perturbation attack with original parameter setting) into the pose transfer learning will extend the training time by more than 900 times, which is not feasible. Thus, we deploy PGD-based attacks [16], including the several variants of fast gradient method (FGM) attacks and PGD attacks, into the framework for the adversarial training. Taking FGM attacks as an example, our final objective function to generate the adversarial samples is defined as below:

$$\begin{aligned} M_{adv\_pose} = M_{clean\_pose} + \\ \epsilon \cdot \text{sign}\left(\nabla f_{adv}(\mathcal{F}(M_{pose}, M_{id}; \theta), M_{GT})])\right), \end{aligned} \tag{7}$$

where $\epsilon$ is the magnitude of the FGM attacks. We encourage readers to refer to the Appendix and [16, 18] for a detailed explanation and implementation of adversarial attacks on 3D data. The training pipeline is demonstrated in Algorithm 1 by taking the FGM attack as an example. With this pipeline, we achieve on-the-fly computation of adversarial samples, which enables the generative model to cover the whole latent pose space via gradients. It's worth emphasizing that our method generates adversarial attacks directly based on the gradients of arbitrary given meshes, with no need to utilize SMPL models in the training.

**Algorithm 1** Adversarial training with FGM for 3D pose transfer.

```
Input: $N$: Total epoch number for training.
  $\lambda$: The threshold to determine a successful attack.
  $\epsilon$: The budget for FGM-based attacks.
  $M_{pose}$: The source pose.
  $M_{id}$: The identity mesh.
  $M_{GT}$: The ground truth mesh.
  $\theta$: The parameter of target model $\mathcal{F}$ to attack.
Output: $\theta'$: The updated parameter of target model $\mathcal{F}$.
  for $epoch$ in $N$ do
    set $\mathcal{F}$ as $eval()$
    $M_{id}.detach(); M_{pose}.detach()$
    $M_{result} = \mathcal{F}(M_{pose}, M_{id}; \theta)$
    $\mathcal{L}_{adv} = f_{adv}(M_{result}, M_{GT})$
    $\mathcal{L}_{adv}.backward()$
    $gradient_{adv} = M_{pose}.grad.detach()$
    $\Delta = gradient_{adv} \times \epsilon$
    $M_{adv\_pose} = M_{pose} + \Delta$
    set $\mathcal{F}$ as $train()$
    $M_{result} = \mathcal{F}(M_{adv\_pose}, M_{id}; \theta)$
    $\mathcal{L}_{rec} = ||M_{result} - M_{GT}||$
    $\mathcal{L}_{rec}.backward()$
  end for
```

## 4. Experiments

In this section, we first present quantitative evaluations of 3D-PoseMAE with two protocols for both clean sample training and adversarial training. One step further, we qualitatively visualize the strong generalization ability of our method as well as the intermediate-generated adversarial samples. Lastly, we perform ablation studies to evaluate the effectiveness of our methods.

**Datasets.** (1) SMPL-NPT [41] is a synthesized dataset containing 24,000 body meshes generated via the SMPL model [4] by random sampling in the parameter space. 16 different identities paired with 400 different poses are provided for training. At the testing stage, 14 new identities are used. Those 400 poses in the training set paired with those new identities will be used as the "seen" protocol and 200 new poses for "unseen" protocols. The models are only trained on the SMPL-NPT dataset and will be generalized to other datasets. (2) FAUST [3] is a well-known 3D human body scan dataset. It provides both FAUST registrations fits the SMPL body model with 6,890 vertices and also the raw

| Methods | PMD ↓ (×$10^{-4}$) | |
|---|---|---|
| | Seen | Unseen |
| DT [19] | 7.3 | 7.2 |
| NPT-MP [41] | 2.1 | 12.7 |
| NPT [41] | 1.1 | 9.3 |
| 3D-CoreNet [37] | 0.8 | - |
| GC-Transformer [10] | **0.6** | 4.0 |
| 3D-PoseMAE (Ours) | **0.6** | **3.4** |

Table 1. Performance comparison with other methods on SMPL-NPT dataset with training on **clean samples**.

scans. We use it for qualitative evaluation. (3) DFAUST [5] dataset contains high-resolution 4D scans of 10 human subjects performing 14 different body motions. We use it for both qualitative and quantitative evaluation.

**Implementation Details.** Our algorithm is implemented in PyTorch [33]. All the experiments are carried out on a server with four Nvidia Volta V100 GPUs with 32 GB of memory and Intel Xeon processors. We train our networks for 400 epochs with a learning rate of 0.00005 and Adam optimizer [24]. The batch size is fixed as 4 for all settings. It takes around 15 hours for pure training (clean samples) on 3D-PoseMAE and 40 hours for adversarial-based training (with FGM attack, which is the fastest) on 3D-PoseMAE. Please refer to the Appendix for more details on parameter settings and other attacks.

The adversarial training of the whole framework is to seek a balance between the magnitude of the adversarial samples and the generative model's robustness. Although it can be controlled by adjusting hyper-parameters such as the attacking budget $\epsilon$, it's intuitive to assume that introducing adversarial samples into the early stage of the training would not contribute as the transferring results are still degenerated and easily fall into local minima and numerical errors in gradient computation. Thus, similar to many existing methods [9, 14] that have trade-offs in training, we conduct the training with two stages. For the first 200 epochs, only clean samples are used to stabilize the model and avoid local minima, and after 200 epochs, the adversarial training starts with adversarial samples added.

### 4.1. Quantitative Evaluation

**SOTA Comparison on Clean Samples.** We start the evaluation with training models on clean samples, following the ***classical evaluation protocol*** for 3D pose transferring from [41] to train the model on the SMPL-NPT dataset. We evaluate the resulting model with Point-wise Mesh Euclidean Distance (PMD) as the evaluation metric:

$$PMD = \frac{1}{|V|}\sum_{\mathbf{v}} \|M_{\mathbf{result}} - M_{\mathbf{GT}}\|_2^2 \,, \tag{8}$$

where $M_{\mathbf{result}}$ and $M_{\mathbf{GT}}$ are the point pairs from the ground truth mesh $M_{\mathbf{GT}}$ and generated one $M_{\mathbf{result}}$. The corresponding experimental results are presented in Table 1.

| Method | PMD ↓ (×$10^{-4}$) | |
|---|---|---|
| | w/o Adversarial Training | w/ Adversarial Training |
| NPT-MP [41] | 307.1 | 62.3 |
| NPT [41] | 237.6 | 59.0 |
| GC-Transformer [10] | 105.2 | 19.3 |
| 3D-PoseMAE (Ours) | **77.6** | **16.9** |

Table 2. Performance comparison with other methods on SMPL-NPT dataset evaluated with **adversarial samples**. Adversarial Training means the models are trained with adversarial samples.

| Datasets | Domains | NPT | 3D-PoseMAE (Ours) |
|---|---|---|---|
| DFAUST [5] | Raw scan | 25.21 | **13.70** |
| SMPL-NPT [41] | Gaussian noise | 12.66 | **6.13** |

Table 3. Quantitative evaluation across datasets. Models are **both** trained on SMPL-NPT in an adversarial manner.

As we can see, our 3D-PoseMAE achieves the SOTA performance with the lowest PMD (×$10^{-4}$) of: 0.6 and 3.4 on "seen" and "unseen" settings. We denote PMD (×$10^{-4}$) as PMD for simplicity in the following. Note that, at this stage, all the methods are trained only with clean samples (no adversarial training/samples involved yet) to make a classical evaluation the same as in previous work.

**SOTA Comparison on Adversarial Samples.** We next conduct the experiments by evaluating models on adversarial samples to verify the performances of models from the robustness aspect. We continually use the PMD as the evaluation metric. To build the testing set of adversarial samples, we directly utilize the "seen" testing set from the SMPL-NPT dataset and generate attack samples (using the same attacking strategy, i.e., the FGM) to attack the victim models. The experimental results are presented in Table 2, where we report the result of using the FGM attack with an attacking budget of 0.08. More experiments with different attack types can be found in the Appendix. When it comes to the pose transferring with adversarial samples, all the methods' performances will suffer from the degeneration problem compared to using the clean pose as the source. Especially for models without adversarial training, the pose transferring performance will drop dramatically, proving the vulnerability of those "clean" models.

**Adversarial Training Improves All the Methods.** To make a fair comparison, we also conducted the adversarial training of two baseline methods, NPT [41] and GC-Transformer [10], with exactly the same setting for the adversarial training pipeline. As we can see in "w/ adversarial training" setting of Table 2, the robustness of all the compared methods has been enhanced considerably. It is worth noting that our 3D-PoseMAE performans the best in both of the strategies, demonstrating it as a strong baseline.

**Evaluation on Noisy Inputs and Raw Scans.** We present extra evaluations on raw scans from the DFAUST dataset and meshes with Gaussian noises on the SMPL-NPT dataset in Table 3. Our 3D-PoseMAE outperforms the compared methods by a large margin.

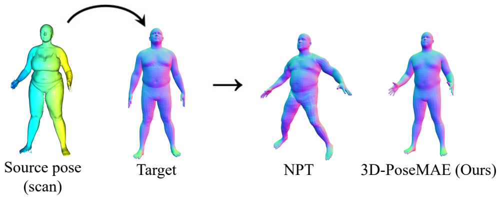

Figure 4. The performance of our method and compared method [41] on an unseen raw scan from the DFAUST dataset [5]. We can see that the compared method failed to handle the raw scan as a source pose, leading to an arbitrary-generated pose while our method can preserve the original pose in a better visual effect.

| Model | Correlation module | Model size | Pose Source | Runtime |
|---|---|---|---|---|
| NPT [41] | - | 24.2M | Mesh | 0.0044s |
| 3D-CoreNet [37] | Correlation matrix | 93.4M | Mesh | 0.0255s |
| GC-Transformer [10] | Spatial-attention | 48.1M | Mesh | 0.0056s |
| 3D-PoseMAE (Ours) | Channel-attention | 40.7M | Mesh/Raw scan | 0.0048s |

Table 4. Model size and runtime of different methods.

**Runtime & Model Size.** We present Table 4 to demonstrate the computational attribute of 3D-PoseMAE. The runtime is obtained by taking the average inference times in the same experimental settings. As shown in the table, the 3D-CoreNet method [37] takes the longest time and largest size compared to other deep learning-based methods. The NPT method [41] has the shortest inference time as there is no correlation module involved thus, the generation performance is degraded. 3D-PoseMAE achieves notable improvements, while the inference time is also encouraging.

## 4.2. Qualitative Evaluation

**Robustness to Various Pose Sources.** Our final goal of introducing adversarial training is to enhance the robustness of the model so that it can be generalized to various noisy and complicated situations. Thus, we also qualitatively display the generality of 3D-PoseMAE over various pose sources, as shown in Fig. 1. Besides, we compare our method with the existing pose transfer method [41] in Fig. 4, to directly transfer pose from raw scans, and the encouraging performance proves that adversarial training can effectively improve the robustness of the models.

**Visualization of Perturbations.** As shown in Fig. 5, we visualize the latent codes and corresponding samples (clean&adversarial). The perturbation causes meaningful pose distribution changes in latent space, even with small magnitude (see 0.0008) that are hard to observe for humans.

**Attacking Effects as By-products.** From Fig. 5, intriguingly, we find that the majority of generated perturbation happened to be located on the body parts that are consistent with the key kinetic positions, such as knees, elbows, and feet, where proven to be easy parts to add attacks.

## 4.3. Ablation Study

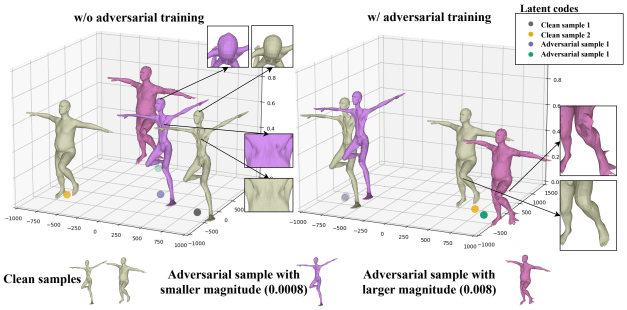

Figure 5. Visualization of latent pose space and the corresponding poses. Please **zoom in for details** due to the page limit.

| Component | Vanilla | + Multi-scale masking | + Channel Attention |
|---|---|---|---|
| PMD ↓ ($\times10^{-4}$) | 4.0 | 3.8 | 3.4 |

Table 5. Ablation study by progressively enabling each component. The rightmost is from the full 3D-PoseMAE.

**Clean vs. Adversarial Training.** We verify the efficacy of adversarial training against the adversarial attacks/noisy inputs in Table 2. We can see that when being attacked by adversarial samples, all the existing methods enjoy benefits from adversarial training compared to the ones being trained with clean samples, e.g., 19.3 vs. 105.2 for GC-Transformer. The results also prove that *generated adversarial samples can successfully attack the pure models*, leading to degenerated results, e.g., the performance of the NPT model degenerates dramatically with PMD from 1.1 (clean pose source) to 237.6 (attacking pose source).

**Each Component.** We verify the contributions made from each component in the 3D-PoseMAE in Table 5. We disable all the key components as a Vanilla model and enable each step by step, showing the contributions can consistently improve the generative performance of the method.

## 5. Conclusion

We work on the robustness problem of the 3D pose transfer, especially on unseen domains and raw noisy inputs from the aspect of adversarial learning. We propose a novel adversarial learning framework customized for 3D pose transfer with a new adversarial function and on-the-fly computation of adversarial samples implementation. We further propose the 3D-PoseMAE with two novelties: a multi-scale masking strategy and a progressive channel-wise attention operation. Experimental results on various data sources show that our method achieves promising performances with substantial robustness to noisy inputs. We show that this work makes end-to-end 3D pose transfer on real-world scans possible.

**Acknowledgment.** This work was supported by the Research Council of Finland (former Academy of Finland) for Academy Professor project EmotionAI (grants 336116, 345122) and ICT 2023 project TrustFace (grant 345948), FARIA (The Finnish-American Research & Innovation Accelerator) project, Infotech Oulu, and the University of Oulu & Research Council of Finland Profi 7 (grant 352788). As well, the authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources.

## References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *TOG*, 39(4): 62–1, 2020. 2, 3

[2] Muhammad Awais, Fengwei Zhou, Hang Xu, Lanqing Hong, Ping Luo, Sung-Ho Bae, and Zhenguo Li. Adversarial robustness for unsupervised domain adaptation. In *ICCV*, pages 8568–8577, 2021. 2, 5

[3] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. Faust: Dataset and evaluation for 3d mesh registration. In *CVPR*, 2014. 1, 6

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2, 6

[5] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *CVPR*, 2017. 1, 7, 8

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 3

[7] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 2, 3, 5

[8] Haoyu Chen, Hao Tang, Nicu Sebe, and Guoying Zhao. Aniformer: Data-driven 3d animation with transformer. *BMVC*, 2021. 1

[9] Haoyu Chen, Hao Tang, Henglin Shi, Wei Peng, Nicu Sebe, and Guoying Zhao. Intrinsic-extrinsic preserved gans for unsupervised 3d pose transfer. In *ICCV*, pages 8630–8639, 2021. 7

[10] Haoyu Chen, Hao Tang, Zitong Yu, Nicu Sebe, and Guoying Zhao. Geometry-contrastive transformer for generalized 3d pose transfer. In *AAAI*, pages 258–266, 2022. 1, 2, 3, 4, 5, 7, 8

[11] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 2023. 1

[12] Haoyu Chen, Hao Tang, Radu Timofte, Luc V Gool, and Guoying Zhao. Lart: Neural correspondence learning with latent regularization transformer for 3d motion transfer. *NeurIPS*, 36, 2024. 1

[13] Jinnan Chen, Chen Li, and Gim Hee Lee. Weakly-supervised 3d pose transfer with keypoints. In *CVPR*, pages 15156–15165, 2023. 3

[14] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodolà. Limp: Learning latent shape representations with metric preservation priors. *ECCV*, 2020. 7

[15] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, pages 2206–2216. PMLR, 2020. 2, 3

[16] Xiaoyi Dong, Dongdong Chen, Hang Zhou, Gang Hua, Weiming Zhang, and Nenghai Yu. Self-robust 3d point recognition via gather-vector guidance. In *CVPR*, pages 11513–11521, 2020. 3, 5, 6

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 3, 6

[19] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 3d-coded: 3d correspondences by deep deformation. In *ECCV*, 2018. 2, 3, 7

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2, 3, 4

[21] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *ICCV*, pages 8988–8999, 2021. 2, 5

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 5

[23] Jaeyeon Kim, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Minimal adversarial examples for deep learning on 3d point clouds. In *ICCV*, pages 7797–7806, 2021. 5, 6

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[25] Hui Kuurila-Zhang, Haoyu Chen, and Guoying Zhao. Adaptive adversarial norm space for efficient adversarial training. *BMVC*, 2023. 2

[26] Kaidong Li, Ziming Zhang, Cuncong Zhong, and Guanghui Wang. Robust structured declarative classifiers for 3d point clouds: Defending adversarial attacks with implicit gradients. In *CVPR*, pages 15294–15304, 2022. 3

[27] Zhouyingcheng Liao, Jimei Yang, Jun Saito, Gerard Pons-Moll, and Yang Zhou. Skeleton-free pose transfer for stylized 3d characters. In *ECCV*. Springer, 2022. 1, 2, 3

[28] Daniel Liu, Ronald Yu, and Hao Su. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *ICIP*, pages 2279–2283. IEEE, 2019. 3, 6

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 2

[30] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *ICCV*, pages 10974–10984, 2021. 1, 3

[31] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 1765–1773, 2017. 5

[32] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. *arXiv preprint arXiv:2203.06604*, 2022. 2, 3, 4

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 7

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 3

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017. 3, 4

[37] Chaoyue Song, Jiacheng Wei, Ruibo Li, Fayao Liu, and Guosheng Lin. 3d pose transfer with correspondence learning and mesh refinement. *NeurIPS*, 34, 2021. 1, 2, 3, 4, 5, 7, 8

[38] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. 5

[39] Jiachen Sun, Weili Nie, Zhiding Yu, Z Morley Mao, and Chaowei Xiao. Pointdp: Diffusion-driven purification against adversarial attacks on 3d point cloud recognition. *arXiv preprint arXiv:2208.09801*, 2022. 3

[40] Tzungyu Tsai, Kaichen Yang, Tsung-Yi Ho, and Yier Jin. Robust adversarial objects against deep learning models. In *AAAI*, pages 954–962, 2020. 2, 3, 5, 6

[41] Jiashun Wang, Chao Wen, Yanwei Fu, Haitao Lin, Tianyun Zou, Xiangyang Xue, and Yinda Zhang. Neural pose transfer by spatially adaptive instance normalization. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[42] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When human pose estimation meets robustness: Adversarial algorithms and benchmarks. In *CVPR*, pages 11855–11864, 2021. 3

[43] Ruibin Wang, Yibo Yang, and Dacheng Tao. Art-point: Improving rotation robustness of point cloud classifiers via adversarial rotation. In *CVPR*, pages 14371–14380, 2022. 5

[44] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *CVPR*, pages 7639–7648, 2021. 2, 3

[45] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *NeurIPS*, pages 2810–2822, 2021. 1, 3

[46] Ziyi Wu, Yueqi Duan, He Wang, Qingnan Fan, and Leonidas J Guibas. If-defense: 3d adversarial point cloud defense via implicit function based restoration. *arXiv preprint arXiv:2010.05272*, 2020. 2, 3, 6

[47] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *CVPR*, pages 9136–9144, 2019. 2, 3, 5, 6

[48] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Gd-mae: generative decoder for mae pre-training on lidar point clouds. In *CVPR*, pages 9403–9414, 2023. 3

[49] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *ICCV*, pages 12498–12507, 2021. 2, 3, 4

[50] Jinlai Zhang, Lyujie Chen, Binbin Liu, Bo Ouyang, Qizhi Xie, Jihong Zhu, Weiming Li, and Yanmei Meng. 3d adversarial attacks beyond point cloud. *arXiv preprint arXiv:2104.12146*, 2021. 2, 3, 5, 6

[51] Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, and Zhigang Tu. Skinned motion retargeting with residual perception of motion semantics & geometry. In *CVPR*, pages 13864–13872, 2023. 2, 3

[52] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022. 2, 3, 4

[53] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *NeurIPS*, 35:27061–27074, 2022. 3

[54] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pages 16259–16268, 2021. 3, 4

[55] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. In *ICCV*, pages 1598–1606, 2019. 5, 6

[56] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 341–357. Springer, 2020. 3

[57] Congcong Zhu, Xiaoqiang Li, Jide Li, and Songmin Dai. Improving robustness of facial landmark detection by defending against adversarial attacks. In *ICCV*, pages 11751–11760, 2021. 5