Geometry-Contrastive Transformer for Generalized 3D Pose Transfer

Haoyu Chen[1], Hao Tang[2], Zitong Yu[1], Nicu Sebe[3], Guoying Zhao[1,*]
[1]CMVS, University of Oulu [2]Computer Vision Lab, ETH Zurich [3]DISI, University of Trento
{chen.haoyu, zitong.yu, guoying.zhao}@oulu.fi
hao.tang@vision.ee.ethz.ch nicu.sebe@unitn.it

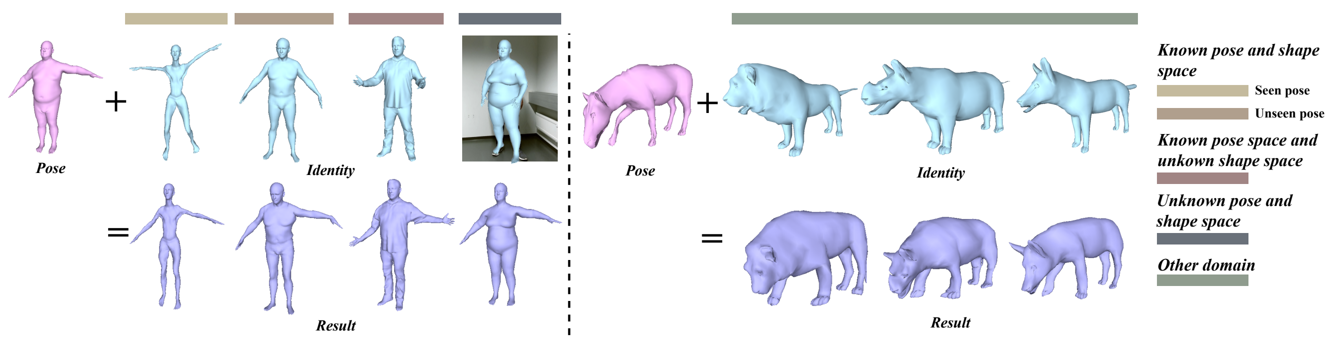

Figure 1: Examples of pose transfer results by our 3D GC-Transformer. Blue, pink, and purple colors stand for identity, pose, and result meshes, respectively. The left part shows the human pose transfer results. The identity meshes are from FAUST (Bogo et al. 2014), MG-cloth (Bhatnagar et al. 2019), SMPL-NPT (Wang et al. 2020), and our new SMG-3D dataset. The right part shows animal pose transfer results on the SMAL dataset (Zuffi et al. 2017). Our method can be generalized to different spaces and even real-world scenarios and animals. More experimental results can be found in supplementary materials.

Abstract

We present a customized 3D mesh Transformer model for the pose transfer task. As the 3D pose transfer essentially is a deformation procedure dependent on the given meshes, the intuition of this work is to perceive the geometric inconsistency between the given meshes with the powerful self-attention mechanism. Specifically, we propose a novel geometry-contrastive Transformer that has an efficient 3D structured perceiving ability to the global geometric inconsistencies across the given meshes. Moreover, locally, a simple yet efficient central geodesic contrastive loss is further proposed to improve the regional geometric-inconsistency learning. At last, we present a latent isometric regularization module together with a novel semi-synthesized dataset for the cross-dataset 3D pose transfer task towards unknown spaces. The massive experimental results prove the efficacy of our approach by showing state-of-the-art quantitative performances on SMPL-NPT, FAUST and our new proposed dataset SMG-3D datasets, as well as promising qualitative results on MG-cloth and SMAL datasets. It's demonstrated that our method can achieve robust 3D pose transfer and be generalized to challenging meshes from unknown spaces on cross-dataset tasks. The code and dataset are made available. Code is available: https://github.com/mikecheninoulu/CGT.

Introduction

Pose transfer, applying the desired pose of a source mesh to a target mesh, is a promising and challenging task in 3D computer vision, which can be widely applied to various industrial fields. However, existing methods (Wang et al. 2020; Cosmo et al. 2020; Zhou, Bhatnagar, and Pons-Moll 2020; Chen et al. 2021b) can only perform well within given datasets of synthesized/known pose and shape space, and fail to be generalized to other unknown spaces with robust performances, which severely limits the further real-world implementations.

To achieve robust performances on unknown latent spaces and other domains as shown in Fig. 1, we propose a novel Transformer network targeting generalized 3D mesh pose transfer. Specifically, a novel geometry-contrastive Transformer with geometrically structured encoders is designed that aims to enhance the identity mesh representation under the guidance of the pose mesh with their *global geometric contrasts*. Locally, we introduce a novel central geodesic contrastive loss to improve the geometric representation by considering the *regional contrast of all the geodesic directions* of each vertex as back-propagation gradients. Furthermore, we present a latent isometric regularization module to stabilize the unreliable performance of cross-dataset pose transfer problems.

Moreover, we present a new 3D mesh dataset, i.e., SMG-3D, for quantitatively evaluating the 3D pose transfer with unknown spaces. The SMG-3D is based on daily spontaneously performed body gestures with more plausible and

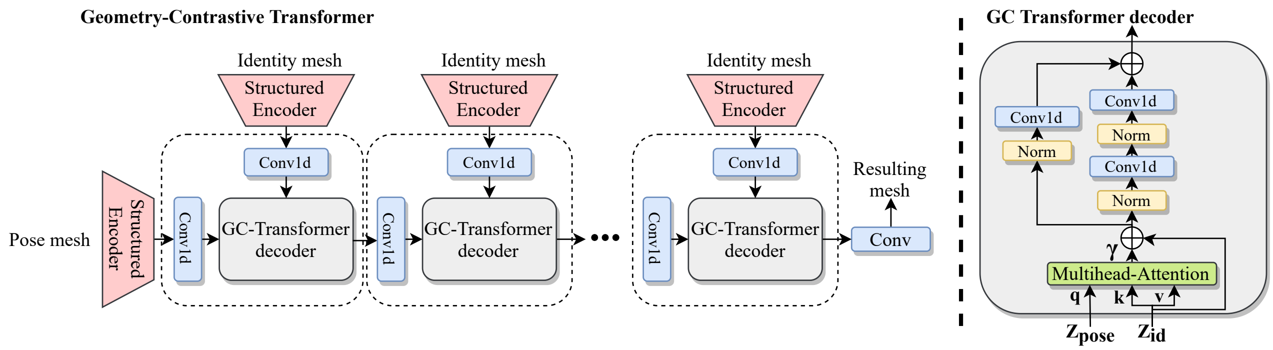

Figure 2: An overlook of our GC-Transformer. The left part is the whole architecture of the GC-Transformer. The right part illustrates the architecture details of one GC-Transformer decoder. The GC-Transformer borrows the idea from the work of (Dosovitskiy et al. 2021) but is extensively extended to 3D data processing tasks for both the encoders and decoders.

challenging body movements and different than those well-performed poses (Mahmood et al. 2019; Bogo et al. 2017). We use a semi-synthesis way to build the dataset to provide necessary GT meshes for training and validating. Our SMG-3D dataset can be jointly combined with other existing body mesh datasets for cross-dataset qualitative analysis.

A natural question to ask is: why not simply use purely synthesized meshes to train and evaluate the model? The short answer is that models trained on purely synthesized meshes will fail in the cross-dataset task. Indeed, using mesh synthesizing models like the SMPL series (Bogo et al. 2016; Zuffi et al. 2017; Pavlakos et al. 2019) can synthesize unlimited poses that can cover the whole latent space, or a large-scale dataset AMASS (Mahmood et al. 2019) to eliminate the inconsistencies with unknown dataset space. However, in practice, even for a small dataset FAUST with only 10 pose categories, it takes more than 26 hours to train a model (Cosmo et al. 2020) to fully learn the latent space. Thus, due to the staggering variability of poses and movements, it's not feasible to train the model with synthesized samples covering the whole pose space. It's desirable that a model can be directly generalized to unknown latent spaces in a more efficient way. To this end, we propose the SMG-3D dataset to tackle the cross-dataset learning issue. It can provide challenging latent distribution allocates on natural and plausible body poses with occlusions and self-contacts instead of well-posed body moves like AMASS (Mahmood et al. 2019), which *could advance the research to real-world scenarios one step further*.

To summarize, our contributions are as follows:

- A novel geometry-contrastive Transformer of positional embedding free architectures with state-of-the-art performances on the challenging 3D pose transfer task.
- A simple and efficient central geodesic contrastive loss that can further improve the geometric learning via preserving the direction gradient of the 3D vertices.
- A challenging 3D human body mesh dataset (i.e., SMG-3D) providing unknown space of naturally plausible body poses with challenging occlusions and self-contacts for the cross-dataset qualitative evaluation.
- A new latent isometric regularization module for adapting to challenging unknown spaces on cross-dataset tasks.

Related work

3D Mesh Deformation Transfer. Deformation transfer aims to generate a new 3D shape with a given pair of source poses and target shapes. Even though existing methods (Groueix et al. 2018; Sumner and Popović 2004) could bring impressive deformation results, the superb performances largely rely on the given correspondences of the source and target meshes, which limits their generalization ability. Some disentanglement-based methods like (Zhou, Bhatnagar, and Pons-Moll 2020; Cosmo et al. 2020; Chen et al. 2021a) tried to decompose meshes into shape and pose factors and achieve pose transfer as a natural consequence. However, extra constraints on the datasets are still needed.

Table 1: A comparison of our GC-Transformer with other 3D Transformer variants.

Model	Vertex operator	Vertex topology	Processed mesh size	Mesh type
Vanilla TFM	MLP	Damaged	-	-
METRO (Lin, Wang, and Liu 2021)	Positional embedding	Preserved, high cost	Down-sampled from 6890 to 431	Pseudo (post-process)
PolyGen (Nash et al. 2020)	Pointer embedding	Preserved, high cost	Filter meshes larger than 800 vertices	Real mesh
GC-Transformer (Ours)	**Depth-wise 1D Conv**	**Preserved, no cost**	**Original size such as 6890**	**Real mesh**

Deep Learning for Geometric Representation. PointNet (Qi et al. 2017a) and PointNet++ (Qi et al. 2017b) have become common-use frameworks that can work directly on sparse and unorganized point clouds. After that, mesh variational autoencoders (Aumentado et al. 2019; Tan et al. 2018) were also proposed to learn mesh embedding for shape synthesis but they are under a strong condition that the shape of target objects should be given as prior. On the other hand, there is a trend to utilize to the self-attention mechanism of Transformers for structural geometric information learning. However, as shown in Table 1, those preliminary works (Lin, Wang, and Liu 2021; Nash et al. 2020; Engel, Belagiannis, and Dietmayer 2020) tried to directly encode the vertex topological structures with computationally demanding embeddings, thus can only handle small-size meshes. In this work, our GC-Transformer is completely different and implements depth-wise 1D Convolution instead of any computational embedding to preserve vertex topological struc-

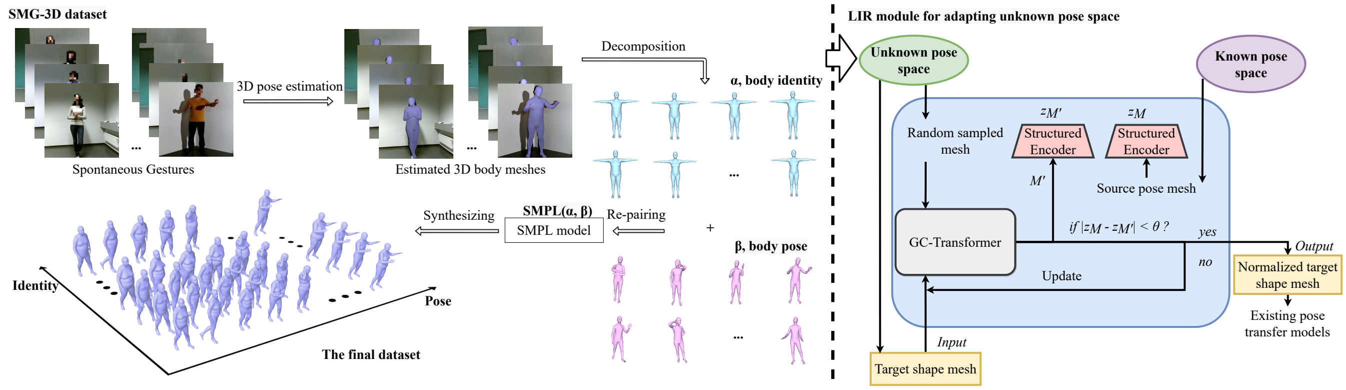

Figure 3: Left: an overlook of our semi-synthesized 3D mesh body gesture dataset SMG-3D. It is a 3D dataset with a pose space that fits the real-world dataset pose distribution, including naturally and spontaneously performed body movements in daily communication with challenging occlusions and self-contacts. Right: the architecture of proposed Latent Isometric Regularization (LIR) module for unknown latent space learning.

tures thus freely handles LARGE meshes with fine-grained details at no cost, which could boost efficient implementations of Transformer frameworks in 3D fields.

Cross-Dataset 3D Pose Transfer. There is few 3D mesh dataset suitable for the pose transfer task. Though many techniques and body models have been developed for 3D data analysis such as SMPL series (Bogo et al. 2016; Romero, Tzionas, and Black 2017; Pavlakos et al. 2019; Zuffi et al. 2017), as well as various 3D human body datasets (Bogo et al. 2014, 2017; Bhatnagar et al. 2019; Pavlakos et al. 2019; Mahmood et al. 2019), they are all originally designed for other tasks such as scan registration, recognition, or shape retrieval. Thus, the poses in those datasets are all exaggerated and perfectly posed actions, for instance, to ensure the quality of the scan registration. However, the latent space distribution of real ones with occlusion and self-contacts can differ widely. Besides, few of the existing datasets can be parameterized and manipulated in the latent space towards desired poses, thus no standard GT is available for the training and the quantitative evaluation. Existing methods (Cosmo et al. 2020) could merely use approximations such as geodesic preservation as substitutes.

Methodology

We define a 3D parametric mesh as $M(\alpha, \beta)$, where α, β denote the parameters of identity (i.e., shape) and pose. Let $M^1(\alpha_{pose}, \beta_{pose})$ be the mesh with the desired pose for style transfer and $M^2(\alpha_{id}, \beta_{id})$ be the mesh with its identity to preserve. Then the polygon mesh $M'(\alpha_{id}, \beta_{pose})$ is the target to generate. The goal of pose transfer is to learn a deformation function f which takes a pair M^1 and M^2 and produces a new mesh M', so that the geodesic preservation of the resulting mesh M' is identical to the source one M^2 and the pose style is identical to M^1.

$$f(M^1(\alpha_{id}, \beta_{id}), M^2(\alpha_{pose}, \beta_{pose})) = M'(\alpha_{id}, \beta_{pose}). \tag{1}$$

Below, we will first introduce how to use the Transformer architecture-based model, called Geometry-Contrastive Transformer (GC-Transformer) for learning the deformation function f, then the Central Geodesic Contrastive (CGC) loss for detailed geometric learning, and at last, the Latent Isometric Regularization (LIR) module for robust pose transfer on cross-dataset tasks.

Geometry-Contrastive Transformer

An overview of the GC-Transformer is depicted in Fig. 2. Our GC-Transformer consists of two key components, one is a structured 3D mesh feature encoder and the other one is a Transformer decoder.

Structured 3D Encoder. As mentioned, existing 3D Transformers needs computationally demanding embeddings to encode vertex positions, thus in practice can only process 'toy' meshes. Inspired by NeuralBody (Peng et al. 2021) that uses structured latent codes to preserve the vertex topology, we modify the conventional PointNet (Qi et al. 2017a) into structured 3D encoders to capture the vertex topology by implementing depth-wise 1D convolution instead of redundant positional embeddings commonly used in conventional Transformers. Meanwhile, we replace the batch normalization layers into Instance Normalization (Ulyanov, Vedaldi, and Lempitsky 2016) layers to preserve the instance style which is widely used on style transfer tasks (Huang and Belongie 2017; Park et al. 2019). The resulting latent embedding vector Z with dimension N_{latent} from the encoder will be dimensionally reduced with 1D convolution and fed into the following GC-Transformer decoder. In this way, LARGE meshes with fine-grained details can be handled freely at no cost by our GC-Transformer while preserving the vertex structures.

GC-Transformer Decoder. We encourage readers to refer to (Dosovitskiy et al. 2021) for a standard Transformer structure, which achieve state-of-the-art results on many tasks such as (Li et al. 2021; Yang et al. 2021). We propose the GC-Transformer decoder that inherits the classical structure with customized designs for 3D meshes. The structure of the GC-Transformer decoder is shown in Fig. 2.

The core difference between the GC-Transformer and a standard Transformer is the design of the multihead self-attention. To learn the correlations between the given meshes

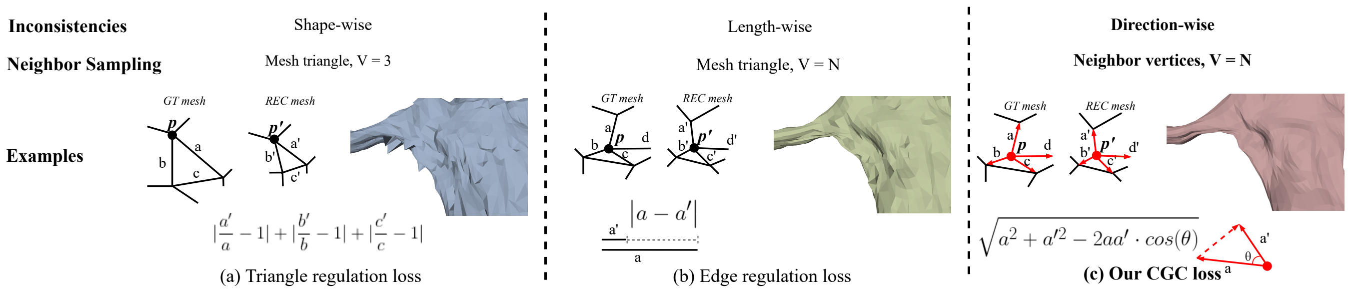

Figure 4: A comparison of different losses for both the neighbor vertex sampling strategy and the local inconsistency. Our CGC loss considers the inconsistencies of all the geodesic directions at each vertex, so that direction gradients can be preserved in the back-propagation. Results show that CGC loss can make the local details more tight and realistic.

for geometric deformation, the model should be able to perceive the geometric information from the two meshes. Thus, we make the inputs of a GC-Transformer as the latent embedding vectors of *two meshes* instead of a single input like the classical Transformer. Besides, as it's a style transfer task, we utilize the Instance Norm introduced by (Huang and Belongie 2017) as our normalization layers. At last, to preserve the structural information of 3D data, the MLP layers are replaced with 1D Convolutional layers.

We denote the latent embedding vectors of the pose mesh and identity mesh from the encoders as Z_{pose} and Z_{id} respectively. We feed the two embedding vectors into different 1D convolution layers to generate the representations $\mathbf{qkv}$ for the standard multihead self-attention (Vaswani et al. 2017). The query $\mathbf{q}$ is from Z_{pose}, and the value $\mathbf{v}$ and key $\mathbf{k}$ are from Z_{id}. Then, the attention weights $A_{i,j}$ based on the geometric pairwise similarity between two elements of $\mathbf{q}$ and $\mathbf{k}$ is given with the following formula:

$$\mathbf{A}_{i,j} = \frac{exp(\mathbf{q}_i \mathbf{k}_j)}{\sum_{i=1}^{n} exp(\mathbf{q}_i \mathbf{k}_j)}. \quad (2)$$

After this, a matrix multiplication between v and the transpose of $\mathbf{A}$ is conducted to perceive the geometric inconsistency between meshes. Finally, we weigh the result with a scale parameter γ and conduct an element-wise sum operation with the original latent embedding Z_{pose} to obtain the refined latent embedding Z'_{pose},

$$Z'_{pose} = \gamma \sum_{i=1}^{n} (\mathbf{A}_{i,j} \mathbf{v}_i) + Z_{pose}, \quad (3)$$

where γ is initialized as 0 and updated gradually during the training with gradients. The obtained Z'_{pose} is followed by typical Transformer operators as introduced above Fig. 2 with a convolutional layer and Tanh activation, generating the final output M'. Please refer to the supplementary materials for more implementing details.

In such a crossing way, the geometric-perceived feature code can consistently be rectified by the original identity mesh and its latent embedding representations. Note that, different than previous attention-based modules (Wang et al. 2018b; Tang et al. 2020b; Huang and Belongie 2017; Tang et al. 2020a), our GC-Transformer could not only compute the pair-wise correlations and contrasts in a crossing-mesh way but also could fully preserve the local geometric details with the residual layer. Most importantly, our GC-Transformer is designed for 3D mesh processing which has never been attempted in these works. Note that input mesh vertices are all shuffled randomly to ensure the network is vertex-order invariant.

Central Geodesic Contrastive Loss

Most of the existing 3D mesh representation learning losses, such as triangle regularization loss, edge loss, Chamfer loss and Laplacian loss (Wang et al. 2018a, 2020; Groueix et al. 2018; Sorkine 2005; Zhou et al. 2020) all repeal the gradient of the direction information of 3D vertices. They only compare the scalar (or weak vector) differences of the mesh vertices such as one-ring geodesic lengths to construct the objective function, while the convexity of the mesh surface containing rich directional gradient information is not utilized. To this end, inspired by the superb performances of central difference convolution (Yu et al. 2020, 2021a,b) that considers the directional difference of depth space, we suggest to compare the vector differences of the vertex topology by proposing a simple yet efficient central geodesic contrastive loss as below:

$$\mathcal{L}_{contra} = \frac{1}{V} \sum_{\mathbf{p}} \sum_{\mathbf{u} \in \Gamma(\mathbf{p})} \sqrt{u_{M'}^2 + u_M^2 - 2u_{M'} u_M \cdot cos(\theta)}, \quad (4)$$

where $\Gamma(\mathbf{p})$ denotes the neighbor edges of vertex $\mathbf{p}$ and V is the total vertex number of the mesh. u_M denotes the edge of mesh M and θ denote the included angle of the edges of u_M and $u_{M'}$. In practice, $\mathcal{L}_{contra}$ can be easily calculated by taking the vector difference of u_M and $u_{M'}$ within the coordinate of each vertex p and divided by the total vertex number as a global normalization.

Our CGC loss has three improvements compared to existing losses: 1) the full inconsistencies of vertex vectors are calculated to preserve the direction gradient; 2) each direction of the vertex is separately considered instead of a simple sum-up; 3) the sampling methods of the neighbor vertices of $\mathbf{p}$ in Eq. (4) is different: the CGC loss samples all the vertices connected to $\mathbf{p}$ resulting in a flexible N neighbor vertices while (Wang et al. 2018a; Groueix et al. 2018) are within the mesh triangle of vertex $\mathbf{p}$ and fixed to 3. Please refer to Fig. 4 for a better understanding. A point-wise $L2$ reconstruction loss of mesh vertices can only capture the absolute distance in the coordinate space. Contrastively, our

CGC loss captures the inconsistencies of all the geodesic directions at each vertex, so that direction gradients can be preserved in the back-propagation. Note that our CGC loss is similar to Laplacian loss but can preserve full vector differences without Laplacian normalization, thus is not only limited to smooth surfaces. As shown in Fig. 4, our CGC loss could offer additional strong supervision especially in tightening the output mesh surface.

Overall Objective Function. With our proposed CGC loss, we define the full objective function as below:

$$\mathcal{L}_{full} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{edge}\mathcal{L}_{edge} + \lambda_{contra}\mathcal{L}_{contra}, \quad (5)$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{edge}$ and $\mathcal{L}_{contra}$ are the three losses used as our full optimization objective, including reconstruction loss, edge loss and our newly proposed CGC loss. λ is the corresponding weight of each loss. In Eq. (5), reconstruction loss $\mathcal{L}_{rec}$ is the point-wise L2 distance and the edge loss (Groueix et al. 2018) is an edge-wise regularization between the GT meshes and predicted meshes.

Cross-Dataset Pose Transfer

Although existing pose transfer methods can deal with fully synthesized/known pose space, they fail to have a robust performance on the pose space that is different from the training one. To facilitate the 3D analysis of human behaviors to real-world implementations, we propose a new SMG-3D dataset as well as a LIR module towards the cross-dataset issue.

A New SMG-3D Dataset. The main contribution of the SMG-3D dataset is providing an alternative benchmark towards cross-dataset tasks by providing standard GTs under a challenging latent pose distribution (unlike perfectly synthesized/performed known distributions). As shown in Fig. 3, SMG-3D is derived from an existing 2D body pose dataset called SMG dataset (Chen et al. 2019) that consists of spontaneously performed body movements with challenging occlusions and self-contacts. Specifically, we first adopt the 3D mesh estimation model STRAPS (Sengupta, Budvytis, and Cipolla 2020) to generate the 3D mesh estimations from the original 2D images of SMG. Then, we select 200 poses and 40 identities as templates to form the potential pose space and optimize them by Vposer (Pavlakos et al. 2019). At last, the generated 3D meshes are decomposed into numerical registrations as latent parameters which are paired to synthesize the resulting 8,000 body meshes via the SMPL model (Bogo et al. 2016), each with 6,890 vertices. Compared to synthesized/well-performed meshes, our in-the-wild 3D body meshes are more practical and challenging with the large diversity and tricky occlusions for providing the unknown latent space. Please check more about our dataset in the supplementary materials.

Latent Isometric Regularization Module. When the poses and shapes are from unknown latent spaces, existing methods suffer from degeneracy in varying degrees (see Fig. 6). We address this issue by introducing the LIR module as shown in Fig. 3 right part, that can aggregate the data distribution of target set and source set. The LIR can be *stacked to existing standard models* to enhance the cross-dataset performance. Specifically, the difference between the two datasets is obtained by comparing the latent pose codes z_M and $z_{M'}$ of the shape mesh M' from the target set and the pose mesh M from the source dataset. The target shape mesh will be fed into GC-Transformer along with another randomly sampled mesh from the target set to obtain a newly generated mesh M'. This will be iteratively executed until the latent pose code difference $z_{M'}$ and z_M converges to less than θ, resulting in a normalized target set. In this way, the latent pose distribution of the target set will be regulated while its isometric information can still be preserved. Essentially, our LIR module serves as a domain adaptive normalization to warm-up the unknown target set to better fit the model trained on the source pose space.

Table 2: Intra-dataset performances on SMG-3D and SMPL-NPT datasets. "NPT MP" stands for NPT model with max pooling layers. Note that the "unseen" setting is still within the same dataset with similar data distributions.

PMD↓ ($\times 10^{-4}$)	Seen			Unseen		
	NPT-MP (Wang et al. 2020)	NPT (Wang et al. 2020)	GC-Transformer	NPT-MP (Wang et al. 2020)	NPT (Wang et al. 2020)	GC-Transformer
SMG-3D	70.3	62.1	**30.7**	120.3	94.6	**52.8**
SMPL-NPT	2.1	1.1	**0.6**	12.7	9.3	**4.0**

Experiments

Datasets

SMPL-NPT (Wang et al. 2020) dataset contains 24,000 synthesized body meshes with the SMPL model (Bogo et al. 2016) by sampling in the parameter space. For training, 16 different identities and 400 different poses are randomly selected and made into pairs as GTs. For testing, 14 new identities are paired with those 400 poses and 200 new poses as "seen" and "unseen" sets. Note that the "unseen" poses are sampled within **the same parameter distribution** as the "seen" poses, thus still in the *same/known latent pose space*.

SMG-3D (Chen et al. 2019) dataset contains 8,000 pairs of naturally plausible body meshes of 40 identities and 200 poses, 35 identities and 180 poses are used as the training set. The rest 5 identities with the 180 poses and the other 20 poses are used for "seen" and "unseen" testing. Note that both SMPL-NPT and SMG-3D provide GT meshes so that they can be used for cross-dataset quantitative evaluation.

FAUST (Bogo et al. 2014) dataset consists of 10 different human subjects, each captured in 10 poses. The FAUST mesh structure is similar to SMPL with 6,890 vertices.

MG-Cloth (Bhatnagar et al. 2019) dataset contain 96 dressed identity meshes with different poses and clothes. The MG-cloth meshes contain way more vertices (above 27,000), which is more challenging for more fine-grained geometry details. Note that meshes in FAUST and MG-cloth are not parameterized SMPL models so geodesic-based approximations (Crane, Weischedel, and Wardetzky 2013) is always used for evaluation in previous works.

SMAL (Zuffi et al. 2017) animal dataset is based on a parametric articulated quadrupedal animal model and we adopted it to synthesize the training and testing datasets.

Intra-Dataset Pose Transfer Evaluation

Firstly, we evaluate the intra-dataset pose transfer performance of our GC-Transformer on the SMPL-NPT and

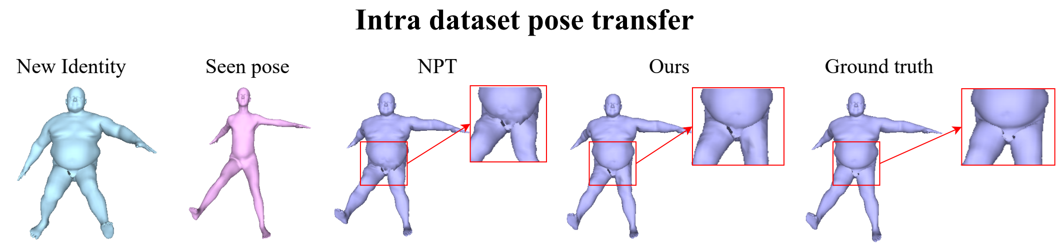

Figure 5: Intra-dataset qualitative results compared with NPT (Wang et al. 2020) on the SMPL-NPT dataset. With satisfying visual effects of both compared methods, our GC-Transformer have a better representation ability in geometry details.

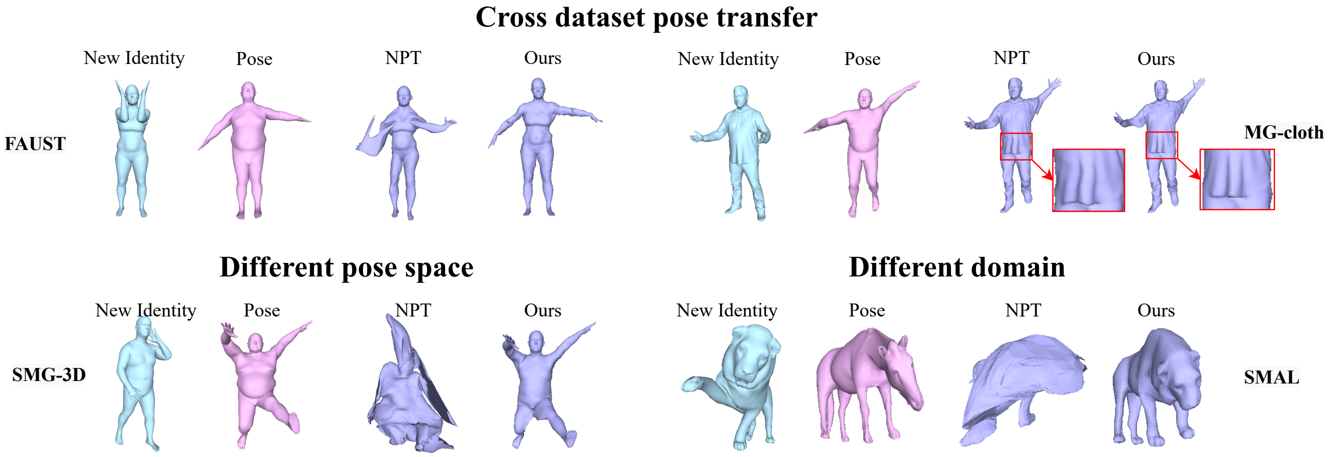

Figure 6: Cross-dataset qualitative results compared with NPT (Wang et al. 2020) on four different datasets. FAUST, SMG-3D and MG-cloth could conduct the pose transfer directly with a model trained on SMPL-NPT.

SMG-3D. Given the GT meshes, we follow (Wang et al. 2020) to adopt Point-wise Mesh Euclidean Distance (PMD) as the evaluation metric:

$$PMD = \frac{1}{|V|} \sum_{\mathbf{v}} \|M_{\mathbf{v}} - M'_{\mathbf{v}}\|_2^2 . \tag{6}$$

where $M_{\mathbf{v}}$ and $M'_{\mathbf{v}}$ are the point pairs from the GT mesh M and generated one M'. The final experimental results can be found in Table 2. For both settings of the SMPL-NPT: "seen" and "unseen pose", our GC-Transformer significantly outperforms compared SOTA methods by more than 45% and 55% with PMD ($\times 10^{-4}$) of: 0.6 and 4.0 vs. 1.1 and 9.3. We denote PMD ($\times 10^{-4}$) as PMD for simplicity in the following. On our SMG-3D dataset, our network again yields the best performance among other methods with PMD of (30.7 and 52.8). As shown, the SMG-3D is more challenging than the SMPL-NPT dataset with way higher PMD values for all the models. Compared to the fully synthesized dataset SMPL-NPT, the poses in SMG-3D are more realistic as they contain many occlusions and self-contacts. The distribution of the poses in the latent space is significantly uneven and discontinuous while the poses synthesized in the SMPL-NPT dataset are way easier with less noise.

Generalized Pose Transfer Evaluation

Cross-Dataset Pose Transfer with Same Pose Space. We extent the setting to cross-datasets by training the model on SMPL-NPT dataset and directly conduct the pose transfer

Table 3: Cross-dataset performances on FAUST dataset. Because we use the raw meshes of FAUST and there is no GT, geometric approximations are used for evaluation.

Disentnaglement Error			
VAE (Aumentado et al. 2019)	LIMP-Euc (Cosmo et al. 2020)	LIMP-Geo (Cosmo et al. 2020)	GC-Transformer
7.16	4.04	3.48	**0.11**

on the unseen meshes from FAUST and MG-cloth datasets. As shown in Fig. 6 first line, NPT might fail when the target pose is not within the training latent space while our method can still perform well. Since there is no GT available here, we adopt the disentanglement error of the pose transfer task illustrated in (Cosmo et al. 2020) as the metrics, see (Cosmo et al. 2020) for more details. In Table 3, we report the performances of GC-Transformer and state-of-the-art models on FAUST. Compared to (Cosmo et al. 2020) trained with the preservation of geodesic distances, ours significantly outperforms (Cosmo et al. 2020): 0.23 vs. 3.48. As expected, the preservation of geodesic distances (Cosmo et al. 2020) can only serve as the approximation of GTs.

Cross-Dataset Pose Transfer with Different Pose Space. In this part, we quantitatively analyze the cross-dataset performance between different latent spaces of SMPL-NPT and SMG-3D datasets by using GTs as metrics. As shown in Table 4. We directly use the model trained on SMPL-NPT to conduct the pose transfer on the meshes from SMG-3D.

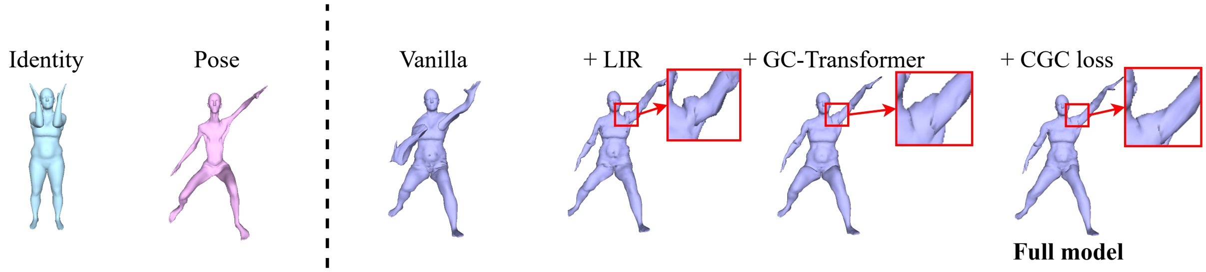

Figure 7: Ablation study by progressively enabling each component. The rightmost mesh is from the full GC-Transformer.

Table 4: Cross-Dataset performances with standard GTs as metrics. Our LIR module can be stacked to existing models and robustly improve the performances on unknown spaces.

Cross-dataset		PMD↓ ($\times 10^{-4}$)		
Training set	Testing set	NPT-MP (Wang et al. 2020)	NPT (Wang et al. 2020)	GC-Transformer
SMPL-NPT	*SMPL-NPT*	*12.7*	*9.3*	*4.0*
	SMG-3D wo/LIR	321.4	240.1	178.7
	SMG-3D w/LIR	132.3	121.4	**79.2**
SMG-3D	*SMG-3D*	*120.3*	*94.6*	*52.8*

The performance of the GC-Transformer (PMD 79.2 and 178.7) keeps outperforming compared methods (PMD 121.4 and 240.1) as presented in Table 4. It can be seen that, by adopting our LIR module, all the models can effectively improve the performances which proves its efficiency, which also proves that the inconsistency of the latent pose space affects the generalization of the pose transfer.

Efficacy of SMG-3D Dataset. From Table 4, we observe that models trained on the synthesized SMPL-NPT dataset can perform well within the same pose space (first row of the table). However, when directly transferring the model to a unknown space like SMG-3D, the PMD dramatically drops down. This proves that a model trained with purely synthesized datasets cannot fit the space distribution of challenging real-world poses. In contrast, by introducing SMG-3D dataset, we can train the model with semi-synthesized data to better fit the pose space of the real-world one, as shown in the last line (PMD improved from 321.4 to 120.3 for NPT and 178.7 to 52.8 for our GC-Transformer). As indicated, a model that works on whole latent pose space is challenging which proves the necessity of our SMG-3D dataset.

Pose Transfer on Different Domain. In the end, we show the robust performance of GC-Transformer on animal pose transfer in Fig. 6. Our model can be directly trained on SMAL dataset without further modification to adapt the non-human meshes, showing a strong generalizing ability.

Ablation Study

Experiments are conducted to present the effectiveness of each proposed component on the SMPL-NPT dataset.

Effect of GC-Transformer. We vary the number of the multi-head attention blocks to show the effect brought by GC-Transformer in Table 5. We observe that the proposed GC-Transformer with four multi-head attention blocks works the best. However, increasing the number of blocks further requires large computational consumption and reaches the GPU memory limits. Thus, we adopt four blocks as default in our experiments.

Table 5: Effect of GC-Transformer. We evaluate the GP-transformer by varying its multihead-attention block number with the rest of the model untouched.

Pose Source	PMD↓ ($\times 10^{-4}$)			
	1 block	2 blocks	3 blocks	4 blocks
Seen-pose	1.4	1.0	0.9	**0.8**
Unseen-pose	7.3	4.9	4.9	**4.2**

Table 6: Effect of CGC loss. We validate the contribution of CGC loss by varying the weight of the CGC loss. As we can see, the CGC loss evidently improves the geometry learning by more than 20%.

Pose Source	PMD↓ ($\times 10^{-4}$)				
	$\lambda_{constra}$=0	0.0005	0.001	0.005	0.05
Seen-pose	0.83	**0.64**	0.84	0.92	1.13
Unseen-pose	4.21	**3.98**	4.27	4.55	4.71

Effect of CGC Loss. We also validate the effect of CGC loss with different $\lambda_{constra}$ settings, as shown in Table 6. It shows that it gains the best performance when $\lambda_{constra}$ is set as 0.0005, which proves that our CGC loss could effectively improve the geometric reconstruction results.

Lastly, we visually present the contributions made from each component in the GC-Transformer in Fig. 7. We disable all the key components as a Vanilla model and enable each step by step. Compared to the Vanilla model, the GC-Transformer, LIR module and CGC loss can consistently improve geometric representation learning. All components can be easily stacked to other existing models.

Conclusion

We introduce the novel GC-Transformer, as well as the CGC loss that can freely conduct robust pose transfer on LARGE meshes at no cost which could be a boost to Transformers in 3D fields. Besides, the SMG-3D dataset together with LIR module can tackle the problem of unstable transferring performance as the cross-dataset benchmark. New SOTA results proves our framework's efficiency in robust and generalized pose transfer. The proposed components can be easily extended to other 3D data processing models.

Acknowledgements

This work was supported by the Academy of Finland for Academy Professor project EmotionAI (grant 336116, 345122) and project MiGA (grant 316765), EU H2020 AI4Media (No. 951911) and Infotech Oulu projects, as well as the CSC-IT Center for Science, Finland, for computational resources.

References

Aumentado, T., Armstrong; Tsogkas, S.; Jepson, A.; and Dickinson, S. 2019. Geometric Disentanglement for Generative Latent Shape Models. In *ICCV*.

Bhatnagar, B. L.; Tiwari, G.; Theobalt, C.; and Pons-Moll, G. 2019. Multi-Garment Net: Learning to Dress 3D People from Images. In *ICCV*.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*.

Bogo, F.; Romero, J.; Loper, M.; and Black, M. J. 2014. FAUST: Dataset and evaluation for 3D mesh registration. In *CVPR*.

Bogo, F.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2017. Dynamic FAUST: Registering human bodies in motion. In *CVPR*.

Chen, H.; Liu, X.; Li, X.; Shi, H.; and Zhao, G. 2019. Analyze Spontaneous Gestures for Emotional Stress State Recognition: A Micro-gesture Dataset and Analysis with Deep Learning. In *FG*.

Chen, H.; Tang, H.; Henglin, S.; Peng, W.; Sebe, N.; and Zhao, G. 2021a. Intrinsic-Extrinsic Preserved GANs for Unsupervised 3D Pose Transfer. In *ICCV*.

Chen, H.; Tang, H.; Sebe, N.; and Zhao, G. 2021b. AniFormer: Data-driven 3D Animation withTransformer. In *BMVC*.

Cosmo, L.; Norelli, A.; Halimi, O.; Kimmel, R.; and Rodolà, E. 2020. LIMP: Learning Latent Shape Representations with Metric Preservation Priors. *ECCV*.

Crane, K.; Weischedel, C.; and Wardetzky, M. 2013. Geodesics in Heat: A New Approach to Computing Distance Based on Heat Flow. *ACM TOG*, 32.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Engel, N.; Belagiannis, V.; and Dietmayer, K. 2020. Point Transformer. *arXiv preprint arXiv:2011.00931*.

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. 3d-coded: 3d correspondences by deep deformation. In *ECCV*.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *ICCV*.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Van Gool, L. 2021. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. *arXiv preprint arXiv:2111.12707*.

Lin, K.; Wang, L.; and Liu, Z. 2021. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*.

Mahmood, N.; Ghorbani, N.; F. Troje, N.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *ICCV*.

Nash, C.; Ganin, Y.; Eslami, S. A.; and Battaglia, P. 2020. Polygen: An autoregressive generative model of 3d meshes. In *ICML*.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 9054–9063.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM TOG*, 36.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2020. Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild. In *BMVC*.

Sorkine, O. 2005. Laplacian mesh processing. In *Eurographics (State of the Art Reports)*.

Sumner, R. W.; and Popović, J. 2004. Deformation transfer for triangle meshes. *ACM TOG*, 23(3): 399–405.

Tan, Q.; Gao, L.; Lai, Y.-K.; and Xia, S. 2018. Variational autoencoders for deforming 3d mesh models. In *CVPR*.

Tang, H.; Bai, S.; Torr, P. H.; and Sebe, N. 2020a. Bipartite graph reasoning gans for person image generation. In *BMVC*.

Tang, H.; Bai, S.; Zhang, L.; Torr, P. H.; and Sebe, N. 2020b. XingGAN for Person Image Generation. In *ECCV*.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, u.; and Polosukhin, I. 2017. Attention is All You Need. In *NeurIPS*.

Wang, J.; Wen, C.; Fu, Y.; Lin, H.; Zou, T.; Xue, X.; and Zhang, Y. 2020. Neural Pose Transfer by Spatially Adaptive Instance Normalization. In *CVPR*.

Wang, N.; Zhang, Y.; Li, Z.; Fu, Y.; Liu, W.; and Jiang, Y.-G. 2018a. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018b. Non-local neural networks. In *CVPR*.

Yang, G.; Tang, H.; Ding, M.; Sebe, N.; and Ricci, E. 2021. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 16269–16279.

Yu, Z.; Wan, J.; Qin, Y.; Li, X.; Li, S. Z.; and Zhao, G. 2021a. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. *IEEE TPAMI*.

Yu, Z.; Zhao, C.; Wang, Z.; Qin, Y.; Su, Z.; Li, X.; Zhou, F.; and Zhao, G. 2020. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, 5295–5305.

Yu, Z.; Zhou, B.; Wan, J.; Wang, P.; Chen, H.; Liu, X.; Li, S. Z.; and Zhao, G. 2021b. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*.

Zhou, K.; Bhatnagar, B. L.; and Pons-Moll, G. 2020. Unsupervised Shape and Pose Disentanglement for 3D Meshes. In *ECCV*.

Zhou, Y.; Wu, C.; Li, Z.; Cao, C.; Ye, Y.; Saragih, J.; Li, H.; and Sheikh, Y. 2020. Fully convolutional mesh autoencoder using efficient spatially varying kernels. In *NerIPS*.

Zuffi, S.; Kanazawa, A.; Jacobs, D. W.; and Black, M. J. 2017. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*.