Adaptive Siamese Masked Autoencoder with Global Optimization for Unsupervised Point Cloud Shape Correspondence

Jiacheng Deng, Jiahao Lu*

University of Science and Technology of China
{dengjc,lujiahao}@mail.ustc.edu.cn

Abstract

Unsupervised point cloud shape correspondence aims to establish point-wise correspondences between point clouds without annotated data. Ensuring efficiency and accuracy is crucial for practically implementing point cloud shape correspondence. Although the current methods have achieved desirable performance, the nature of encoding at dense points limits their application in actual scenarios. Moreover, independently computing per-point correspondences results in numerous multiple-to-one erroneous correspondences. To address these issues, we present an *Adaptive siamese Masked autoencoder wIth Global Optimization* (AMIGO), comprising a siamese masked autoencoder and a global optimization module. In the siamese masked autoencoder, we downsample the input point cloud and employ adaptive siamese mask operations to boost the coding capabilities of the encoder, thereby mitigating the information loss caused by downsampling. In the global optimization module, optimal transport is only utilized to generate pseudo-labels during the training phase, facilitating the efficient global planning of the correspondence results. Extensive experiments on four standard human and animal benchmarks demonstrate that AMIGO surpasses existing methods with remarkable margins, achieving new state-of-the-art results.

Introduction

As one of the most indispensable techniques for real-world applications like mixed reality (Mahmood and Han 2019; Lachat, Landes, and Grussenmeyer 2019), shape editing (Liu et al. 2019a), and non-rigid human body alignment (Brown and Rusinkiewicz 2007), point cloud shape correspondence aims to establish precise point-to-point correspondences between source and target point clouds. However, non-rigid objects exhibit shape variations and body part similarity, posing significant challenges in efficiently and accurately establishing point-to-point correspondences.

To tackle the aforementioned challenges, a series of point cloud shape correspondence methods has been introduced, broadly classified into spectral-based (Bronstein, Bronstein, and Kimmel 2006; Huang et al. 2008; Tevs et al. 2011) and point-based (Deprelle et al. 2019; Zeng et al. 2021; Lang et al. 2021; Deng et al. 2023; He et al. 2023). Spectral-based approaches calculate functional mappings between projected features and learn transformations to establish correspondences among deformable shapes, proven effective for mesh data. Nevertheless, these methods are limited by complex preprocessing and reliance on additional vertex connectivity information. For point-based methods, fully supervised strategies (Deprelle et al. 2019; Groueix et al. 2018; Marin et al. 2020) use annotated data to guide point-wise correspondences in point cloud pairs, obviating extra requirements for connectivity information. However, fully supervised methods still heavily rely on time-consuming point-wise annotations. To ease annotation burden, researchers are increasingly exploring unsupervised point cloud shape correspondence tasks (Zeng et al. 2021; Lang et al. 2021; Deng et al. 2023; He et al. 2023). Current methods target specific challenges in point cloud shape correspondence, addressing issues like rotation sym-

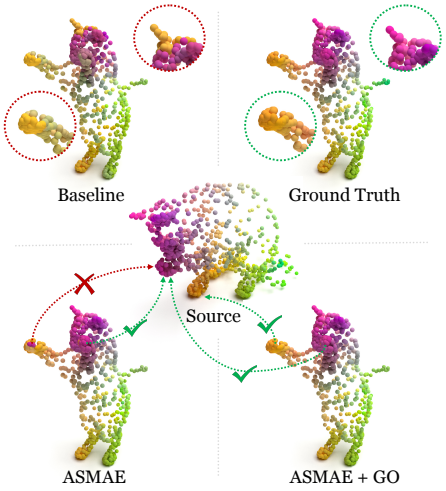

Figure 1: **The visualization of dense point matching results.** The correspondences are visualized by transferring colors from source to target according to matching results. ASMAE corrects most of the mismatches in the baseline, such as cat ears and front limbs. GO further refines the multiple-to-one erroneous matches duiring training phase.

*Corresponding Author

metry (Deng et al. 2023) and shape differences (He et al. 2023) to achieve substantial accuracy improvements without relying on annotated data or connectivity information. However, these methods share a standard paradigm, where all 1024 points from the source and target point clouds are fed into an encoder for point feature extraction. Subsequently, the shape correspondences are established based on point-to-point feature similarity. The simultaneous encoding of numerous points severely limits inference speed, hindering real-time applicability. Moreover, pursuing the optimal match for each point independently through pointwise similarity is a suboptimal strategy, as it may lead to multiple-to-one matches. As illustrated in Figure 1, the forelimb and ear are mistakenly matched to a hindlimb. Thus, the optimal strategy requires a global planning of correspondences.

By analyzing previous point-based methods, two critical issues are identified for accurate correspondence: 1) *How to accelerate the network inference for point cloud shape correspondence?* Downsampling is widely adopted to accelerate inference across various 3D tasks (Wang et al. 2023; Schult et al. 2022; Lu et al. 2023; Li et al. 2024; Deng, Lu, and Zhang 2025). However, downsampling inadvertently reduces the density of point clouds, leading to a loss of local details and an oversimplification of the overall body structure. The current paradigms in the point cloud shape correspondence field lack the eligible capability to provide sufficiently elaborate structural representations for sparse point clouds. The Masked Autoencoder (MAE) (He et al. 2022) promises improved encoding of fine-grained structures within the encoder. Nonetheless, directly applying MAE to source and target point clouds without interaction is insufficient to encode features that possess both intra-discriminative and inter-correspondence qualities. Besides, the random masks used in MAE do not adequately support discriminative representations. Therefore, it is imperative to design an efficient paradigm to accelerate inference while adeptly ensuring intra-discriminative and inter-correspondence modeling. 2) *How to effectively mitigate multiple-to-one erroneous correspondences?* Calculating the optimal match for each point individually may have ambiguous activations without pointwise annotations, often resulting in multiple-to-one erroneous matches in Figure 1. Thus, an advanced matching strategy should eliminate the convention of pursuing an optimal match for each point independently and turn to the ideology of global optimum, in other words, finding the global high confidence match for all points in a point cloud pair.

To achieve the above goal, we propose an *Adaptive Siamese Masked AutoEncoder with Global Optimization* (AMIGO) for unsupervised point cloud shape correspondence, including an adaptive siamese masked autoencoder (ASMAE) and a global optimization module (GO). In the adaptive siamese masked autoencoder, the point cloud pair interacts via masking partial source point cloud structures and recovering them based on the complete target point cloud (as illustrated in Figure 4). This interaction promotes better correspondences, enabling intra-discriminative and inter-correspondence modeling on downsampled point clouds. Besides, an Adaptive Adversarial Mask Generator (AAMG) is proposed to generate point cloud masks in the source point cloud adaptively based on specific shapes. Despite an eightfold downsampling of the point cloud, our ASMAE achieves more accurate correspondences. In the global optimization module, optimal transport (Cuturi 2013) assigns pseudo-labels to each point through a global perspective in refinement for correspondence optimization. The global optimum effectively mitigates multiple-to-one erroneous correspondences. Notably, We only use the optimization to get pseudo labels during the training phase.

In summary, our contributions are outlined as follows: (i) We innovatively develop an adaptive siamese masked autoencoder with global optimization for unsupervised point cloud shape correspondence, to the best of our knowledge, which is the first point cloud siamese masked autoencoder. (ii) We propose an adaptive siamese masked autoencoder that accelerates network inference and improves correspondence accuracy. Furthermore, the global optimization module effectively alleviates the multiple-to-one correspondence issue. (iii) Extensive experiments on four standard benchmarks, including SURREAL (Groueix et al. 2018), SHREC'19 (Melzi et al. 2019), SMAL (Zuffi et al. 2017), and TOSCA (Bronstein, Bronstein, and Kimmel 2008), demonstrate AMIGO outperforms current sota methods.

Related Work

Point Cloud Representation Learning. Early deep learning research (Poux and Billen 2019; Liu et al. 2019b; Borges, Garcia, and de Queiroz 2022) adapt CNNs for point cloud data by projecting onto 2D bird's-eye view or 3D volumetric grids to address representation challenges. To overcome voxelization's quantization error, PointNet (Qi et al. 2017a) extracts pointwise features using multi-layer perceptrons and max pooling. PointNet++(Qi et al. 2017b) improves local information integration through a hierarchical sampling and grouping approach. DGCNN(Wang et al. 2019) utilizes the point cloud's graph-like structure to learn the edge features of the graph in each layer. Moreover, attention-based techniques (Vaswani et al. 2017; Liu et al. 2021; Guo et al. 2021; Han et al. 2022) are gaining momentum. Point Transformer (Zhao et al. 2021) utilizes subtraction vector attention for local features. In contrast, PCT (Guo et al. 2021) is a global attention network.

Point Cloud Masked Autoencoders. Masked autoencoders have demonstrated remarkable performance in self-supervised learning for point clouds. Point-MAE (Pang et al. 2022) pioneers MAE-style pre-training, proving competitive across various tasks. Point-M2AE (Zhang et al. 2022) uses a hierarchical approach for multi-scale spatial geometry understanding. Point2vec (Abou Zeid et al. 2023) implicitly reconstructs masked patches, removing the need for intricate distance metrics. Recently, I2P-MAE (Zhang et al. 2023) explores 2D pre-trained models for 3D mask generation. However, the aforementioned methods all deal with individual point clouds. Therefore, we introduce the first point cloud siamese masked autoencoder explicitly designed for point cloud pairs, enhancing interaction between the point cloud pair with an innovative adversarial masking strategy.

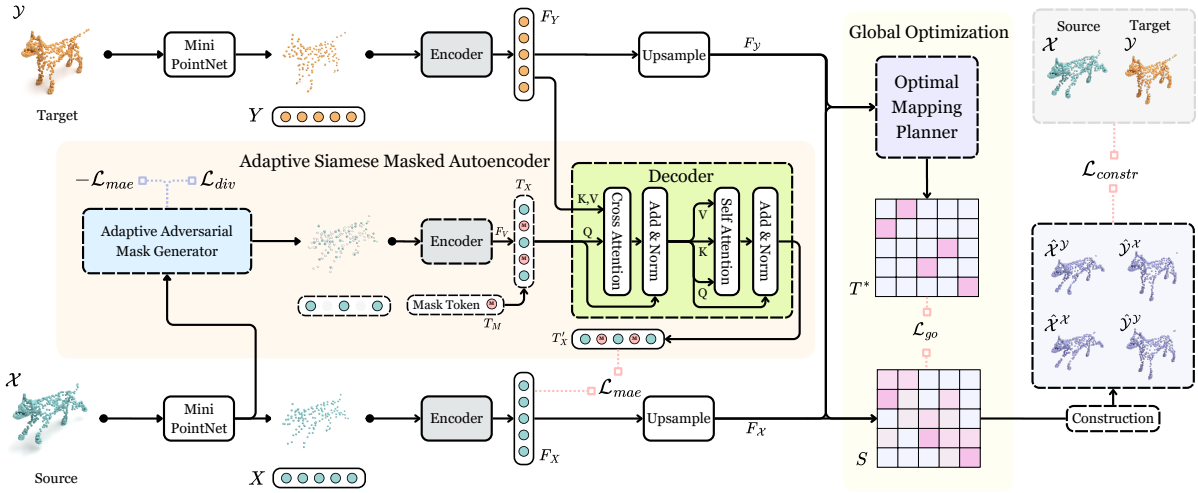

Figure 2: **Illustration of the AMIGO.** AMIGO mainly consists of the Adaptive Siamese Masked Autoencoder and the Global Optimization. The Adaptive Siamese Masked Autoencoder masks and recovers the source point cloud via the complete structure of the target point cloud, enhancing correspondence perception capabilities. Additionally, the global optimization generate pseudo-labels in a global view, facilitating efficient optimization of similarity relationships during training phase.

Shape Correspondence. Point cloud shape correspondence methods can be broadly categorized into spectral-based methods and point-based methods. Spectral-based methods (Bronstein, Bronstein, and Kimmel 2006; Huang et al. 2008; Tevs et al. 2011; Roufosse, Sharma, and Ovsjanikov 2019; Donati, Sharma, and Ovsjanikov 2020) employ mesh data, which includes vertex coordinates and connectivity information. Traditional methods (Bronstein, Bronstein, and Kimmel 2006; Huang et al. 2008; Tevs et al. 2011) compute the Laplacian-Beltrami Operator (LBO) eigenvectors as basis functions and derive a linear transformation for shape correspondence. Cao et al. (Cao and Bernard 2023) transfer rich structural information from mesh data to point clouds. Point-based methods (Corman, Ovsjanikov, and Chambolle 2014; Zeng et al. 2021; Lang et al. 2021; He et al. 2023; Deng et al. 2023; Deng, Lu, and Zhang 2024; Dutt, Muralikrishnan, and Mitra 2024) directly process point clouds without connectivity information. CorrNet3D (Zeng et al. 2021) and DPC (Lang et al. 2021) use DGCNN (Wang et al. 2019) network to aggregate features via graph structure. HSTR (He et al. 2023) develops a multi-receptive-field transformer point cloud encoder that considers local structures and global semantics. SE-ORNet (Deng et al. 2023) incorporates an orientation estimation module to solve the rotational symmetry problem. Nevertheless, the inference speed is limited by encoding thousands of points concurrently. To address the limitation, we introduce a point cloud Siamese MAE to represent downsampled point clouds efficiently and yield more accurate correspondences.

Method

Overview

Unsupervised point cloud shape correspondence aims to establish a one-to-one mapping, $f : \mathcal{X} \rightarrow \mathcal{Y}$, between two point clouds (source $\mathcal{X}$ and target $\mathcal{Y}$) without any annotations. Figure 2 illustrates our approach's pipeline, including two key components: an adaptive siamese masked autoencoder and a global optimization module. Given the source point cloud $\mathcal{X} \in \mathbb{R}^{N\times 3}$ and the target point cloud $\mathcal{Y} \in \mathbb{R}^{N\times 3}$, we initially perform eightfold downsampling using a mini PointNet (Qi et al. 2017a) to reduce and cluster the points, resulting in $X \in \mathbb{R}^{N_d\times C_1}$ and $Y \in \mathbb{R}^{N_d\times C_1}$. Subsequently, X and Y undergo feature aggregation via the shared encoder, resulting in F_X and F_Y. To enhance the intra-discriminative and inter-correspondence characteristics of the features, we design the Adaptive Siamese MAE. Further details on this aspect will be discussed in Section . F_X and F_Y undergo upsampling and interpolation to obtain the complete representations, $F_{\mathcal{X}} \in \mathbb{R}^{N\times C}$ and $F_{\mathcal{Y}} \in \mathbb{R}^{N\times C}$, and to compute the similarity S and construction loss $\mathcal{L}_{constr}$ between the source and target point clouds. Furthermore, $F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$ are also introduced into the global optimization module. In this phase, optimal mapping as $\mathbf{T}^*$ is derived through optimal transportation, which guides the refinement of S via $\mathcal{L}_{go}$. This aspect will be elaborated upon in detail in Section . Finally, we introduce the model training and inference under the unsupervised setting in Section .

Adaptive Siamese Masked Autoencoder

To improve the encoder's ability to capture local structural information in sparse point clouds and enhance correspondence, we introduce the adaptive siamese masked autoencoder, consisting of the adaptive adversarial mask generator and the siamese masked autoencoder.

Adaptive adversarial mask generator. The adaptive adversarial mask generator employs adversarial learning strategies to adaptively generate masks for key structural points within a point cloud based on its specific shape. This approach aids in establishing intra-discriminative representations for point clouds by increasing the challenge of reconstructing the complete point cloud structure through the siamese masked autoencoder. As shown in Figure 3,

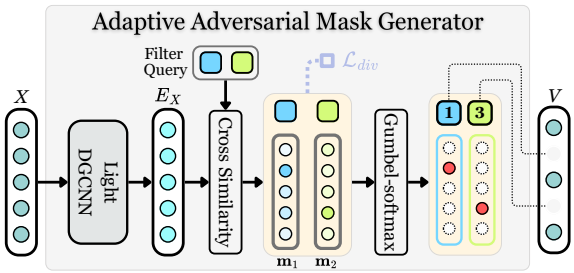

Figure 3: **Mask generator structure.** The predicted point positions in the output are constrained to match the true point positions.

$X \in \mathbb{R}^{N_d \times C_1}$ is fed into a light DGCNN (Wang et al. 2019) to extract features among sparse point clouds, resulting in $E_X \in \mathbb{R}^{N_d \times C_2}$. A set of N_m learnable filter queries$\in \mathbb{R}^{N_m \times C_2}$ are established to adaptively select points to mask from various perspectives. Specifically, each filter query computes a similarity vector $\mathbf{m}^x$ with E_X and employs Gumbel-softmax (Jang, Gu, and Poole 2016) to select the most similar point in X as the masked point. Therefore, N_m filter queries can generate N_m correspond masked points. To promote masked points mutually exclusive and almost evenly distributed, we introduce a diversity loss $\mathcal{L}_{\text{div}}$ that operates on similarity vectors:

$$\mathcal{L}^X_{\text{div}} = \frac{1}{{N_m}^2} \sum_{i=1}^{N_m} \sum_{j=1}^{N_m} \frac{\mathbf{m}^x_i \mathbf{m}^x_j}{||\mathbf{m}^x_i||_2 ||\mathbf{m}^x_j||_2}, \tag{1}$$

where $\mathbf{m}^x_i$ represents the similarity vector between the ith filter query and E_X. However, occasional duplicate masked points may lead to different number of mask points for different samples within a batch, impacting parallel computation. In practice, we randomly select the extra points to ensure the total number of masked points equals N_m.

It is worth noting that the training of the Adaptive Adversarial Mask Generator is alternately updated with the other network modules. The loss function employed for this module steers the module towards generating challenging mask. Specifically, the adversarial loss can be formulated as:

$$\mathcal{L}^X_{mask} = \mathcal{L}^X_{div} - \mathcal{L}^X_{mae}, \tag{2}$$

where $\mathcal{L}^X_{mae}$ denotes the training loss function in the siamese masked autoencoder, explained in the following.

Siamese masked autoencoder. Using point cloud masks generated by the adaptive adversarial mask generator, we feed the visible point cloud into a point cloud encoder to obtain point cloud features, denoted as $F_V \in \mathbb{R}^{(N_d - N_m) \times C}$. For the encoder, we employ the local instead of global self-attention, similar to HSTR (He et al. 2023), to constitute the transformer-based encoder. For the decoder, we utilize standard point cloud cross-attention blocks and self-attention blocks.

In the subsequent process of using the decoder to reconstruct the masked regions, we do not employ the traditional MAE method, which is applied independently to individual entities. Instead, we utilize the features of the target point cloud to restore the masked regions of the source point cloud. This approach facilitates the identification of corresponding structures between the source and target point clouds. Concretely, a learnable mask token $T_M \in \mathbb{R}^C$ is introduced and duplicated to supplement the missing mask point tokens, resulting in T_X. T_X is utilized as a query, and the target point cloud representation F_Y serves as both key and value. These components are jointly input into the decoder, resulting in the production of an updated T'_X. As illustrated in Figure 4, the decoding process plays a crucial role in enhancing the model's correspondence and construction capabilities. The updated T'_X constructs the source point cloud by leveraging the feature similarity S_X between T'_X and F_X, along with the downsampled source point coordinates $P_X \in \mathbb{R}^{N_d \times 3}$, as follows:

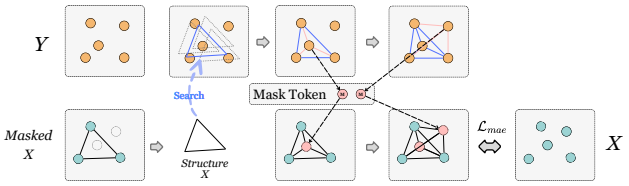

Figure 4: **Autoencoder mechanism diagram.** The masked X point cloud initially builds its visible structure in the encoder and then locates the corresponding structure within the Y. By modeling the geometric relationship in the Y, the mask tokens infer the approximate location of the masked points in the Masked X based on the established structural relationship. The recoverd X structure undergoes internal adjustments, and the cross-construction loss (L_{mae}) is designed to align the true structure and the recoverd stucture.

$$\hat{X}_{t_i} = \sum_{j \in \mathcal{N}_{F_X}(t_i)} \frac{e^{s_{ij}}}{\sum_{l \in \mathcal{N}_{F_X}(t_i)} e^{s_{il}}} x_j, \tag{3}$$

where $t_i \in T'_X$, $s_{ij} \in S_X$ and $x_j \in P_X$. $\mathcal{N}_{F_X}(t_i)$ represent the k-nearest neighbors of t_i in the target feature space F_X. The cross-construction of the source point cloud P_X by the T'_X is denoted as $\hat{X}_T \in \mathbb{R}^{N_d \times 3}$, where $\hat{X}^i_T = \hat{X}_{t_i}$. Then, we constrain the training with the mae loss as follows:

$$\mathcal{L}^X_{mae} = \mathbf{CD}(P_X, \hat{X}_T), \tag{4}$$

where $\mathbf{CD}$ means the chamfer distance.

Symmetric training. In the unsupervised point cloud shape correspondence task, the statuses of source and target point clouds are interchangeable. Therefore, during the training of ASMAE, we can leverage symmetric training on point cloud pairs. Specifically, in addition to computing $\mathcal{L}^X_{mae}$ and $\mathcal{L}^X_{mask}$ as described above, we treat the target point cloud as the source and vice versa. Consequently, we calculate $\mathcal{L}^Y_{mae}$ and $\mathcal{L}^Y_{mask}$ using the same formulas. The specific computation formulas are as follows:

$$\begin{aligned} \mathcal{L}^Y_{mae} &= \mathbf{CD}(P_Y, \hat{Y}_T), \\ \mathcal{L}^Y_{mask} &= \mathcal{L}^Y_{div} - \mathcal{L}^Y_{mae}. \end{aligned} \tag{5}$$

The final loss for the siamese masked autoencoder is denoted as $\mathcal{L}_{mae}$, and for the adaptive adversarial mask generator, it

is $\mathcal{L}_{mask}$:

$$\begin{aligned}\mathcal{L}_{mae} &= \frac{1}{2}(\mathcal{L}^X_{mae} + \mathcal{L}^Y_{mae}),\\ \mathcal{L}_{mask} &= \frac{1}{2}(\mathcal{L}^X_{mask} + \mathcal{L}^Y_{mask}).\end{aligned} \tag{6}$$

Global Optimization

The Global Optimization is crafted to assign pseudo-labels to each point during training, providing a global refinement . Specifically, F_X and F_Y are upsampled to obtain complete point features $F_{\mathcal{X}}$ and $F_{\mathcal{Y}}$. Subsequently, the similarity matrix S is calculated via the cosine similarity. During inference, point-to-point correspondences between X and Y are established by assigning the nearest point y_{j^*} in the feature space to each point x_i in X as its corresponding point. This selection rule can be expressed as follows:

$$f\left(x_i\right) = y_{j^*}, j^* = \underset{j}{\operatorname{argmax}}(s_{ij}). \tag{7}$$

Without the assistance of point-level GT annotations, the learned similarity matrix may have ambiguous activations. To refine the similarity matrix, we propose to assign each point a pseudo correspondence label and find the global high confidence match for all points in a point cloud pair. The pseudo label assignment problem is the same as the Optimal Transport (OT), and the goal of the OT problem is to find a transportation plan $\mathbf{T}^*$ at a global minimal transportation cost, which can be estimated by the Sinkhorn algorithm with linear programming (Cuturi 2013). Given the target point representations $F_{\mathcal{Y}} \in \mathbb{R}^{N\times C}$, we are interested in mapping the source point representations $F_{\mathcal{X}} \in \mathbb{R}^{N\times C}$ to the target points. We denote this mapping as $\mathbf{T}^*$ with a similarity cost matrix $(1 - F_{\mathcal{Y}}F_{\mathcal{X}}^{\top})$. The optimal transportation plan $\mathbf{T}^* \in \mathbb{R}^{N\times N}$ is obtained by minimizing the similarity cost with a regularization term:

$$\mathbf{T}^* = \min_{\mathbf{T}\in\mathcal{T}} \mathrm{Tr}(\mathbf{T}^{\top}(1 - \mathrm{F}_{\mathcal{Y}}\mathrm{F}_{\mathcal{X}}^{\top})) - \epsilon \mathrm{H}(\mathcal{T}), \tag{8}$$

where $H(\mathcal{T}) = -\sum_{ij} \mathbf{T}_{ij}\mathrm{log}\mathbf{T}_{\mathrm{ij}}$ is the entropy function. $\mathcal{T}$ is the search space. Following the works (Caron et al. 2020; Ge et al. 2021), $\mathcal{T}$ is limited as $\mathcal{T} = \{\mathbf{T} \in \mathbb{R}^{N\times N} | \mathbf{T}\mathbb{1} = \frac{1}{N}\cdot\mathbb{1}, \mathbf{T}^{\top}\mathbb{1} = \frac{1}{N}\cdot\mathbb{1}\}$, where $\mathbb{1}$ denotes the vector of all ones. The problem stated in Eq. 8 can be efficiently estimated using the Sinkhorn Iteration (Cuturi 2013) on a GPU. To refine the similarity matrix, the pseudo label loss is imposed to make the similarity matrix S approach $\mathbf{T}^*$:

$$\mathcal{L}_{go} = \mathbf{CE}(\mathbf{T}^*, S), \tag{9}$$

where $\mathbf{CE}$ is the Cross-Entropy loss. Notably, we use the Sinkhorn algorithm only during training to generate pseudo-labels that guide model optimization, thereby avoiding any computational overhead during inference.

Model Training & Inference

Following previous point cloud shape correspondence methods (Lang et al. 2021; Deng et al. 2023; He et al. 2023), we also use construction losses $\mathcal{L}_{constr}$ to enhance feature smoothness and discrimination. The specific computational details will be elaborated in the supplementary materials. Eventually, the total loss in our approach is composed of three key components: the construction loss for point cloud pairs ($\mathcal{L}_{constr}$), the cross-construction loss within the ASMAE ($\mathcal{L}_{mae}$), and the global optimization loss ($\mathcal{L}_{go}$), formulated as:

$$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{constr} + \lambda_2 \cdot \mathcal{L}_{mae} + \lambda_3 \cdot \mathcal{L}_{go}, \tag{10}$$

where λ_i are hyperparameters, balancing the contribution of different loss terms. The training loss for the adaptive adversarial mask generator is $\mathcal{L}_{mask}$. During inference, the adaptive siamese masked autoencoder and global optimization are no longer computed. The downsampled point cloud undergoes encoding to aggregate sparse point cloud features. Subsequently, after upsampling to obtain complete point cloud features, point-wise similarity computation is performed to establish correspondences.

Experiment

Experimental Setup

Datasets. To validate the effectiveness and generalization of our method, we conducted experiments following the established protocol (Lang et al. 2021; He et al. 2023) on human datasets (SURREAL (Groueix et al. 2018) and SHREC'19 (Melzi et al. 2019)) as well as animal datasets (SMAL (Zuffi et al. 2017) and TOSCA (Bronstein, Bronstein, and Kimmel 2008)). For SURREAL, containing 230,000 training shapes, we randomly downsample to 2000 shape pairs. SHREC'19, with 44 real human models, is paired into 430 examples. For the animal datasets, the large-scale SMAL dataset are parameterized models. TOSCA, generated by deforming templates (human, dog, and horse), yields a training set of 260 and a test set of 286 samples among 41 figures. To further validate the robustness and generalization, experiments are also conducted on the real-scanned Owlii dataset (Xu, Lu, and Wen 2017) and the partial-shape SHREC'16 dataset (Cosmo et al. 2016). A detailed discussion can be found in the Robustness Analysis section of the supplementary materials.

Evaluation metrics. The evaluation metrics encompass average correspondence error and correspondence accuracy. The average correspondence error, based on the Euclidean measure for a source and target point cloud pair (X, Y):

$$err = \frac{1}{n}\sum_{x_i\in X} \left\| f\left(x_i\right) - y_{gt} \right\|_2, \tag{11}$$

where $y_{gt} \in Y$ is the ground-truth matching point to x_i. The correspondence accuracy is formulated as:

$$acc(\epsilon) = \frac{1}{n}\sum_{x_i\in X} \mathbb{1}\left(\left\| f\left(x_i\right) - y_{gt} \right\|_2 < \epsilon d\right), \tag{12}$$

where $\mathbb{1}(\cdot)$ is the indicator function, d is the maximum Euclidean distance between points in Y, and $\epsilon \in [0, 1]$ denotes the error tolerance, usually set to 0.01.

Implementation details. The mini PointNet achieves an $8\times$ downsampling, generating a sparse point cloud with 128 points by employing two layers of simple PointNet structures. This sparse input is then fed into a Siamese masked

Method	Data	SHREC'19 (SHREC'19)		SHREC'19 (SURREAL)		TOSCA (TOSCA)		TOSCA (SMAL)	
		acc ↑	err ↓	acc ↑	err ↓	acc ↑	err ↓	acc ↑	err ↓
SURFMNet	Mesh	5.9%	0.2	4.3%	0.3	/	/	/	/
GeoFMNet	Mesh	/	/	8.2%	0.2	/	/	/	/
Diff-FMaps	Point	/	/	4.0%	7.1	/	/	/	/
3D-CODED	Point	/	/	2.1%	8.1	/	/	0.5%	19.2
Elementary	Point	/	/	2.3%	7.6	/	/	0.5%	13.7
CorrNet3D	Point	0.4%	33.8	6.0%	6.9	0.3%	32.7	5.3%	9.8
DPC	Point	15.3%	5.6	17.7%	6.1	34.7%	2.8	33.2%	5.8
SE-ORNet	Point	17.5%	5.1	21.5%	4.6	38.3%	2.7	36.4%	3.9
HSTR	Point	19.3%	4.9	19.4%	5.6	52.3%	1.2	33.9%	5.6
TANet	Point	21.5%	4.5	20.6%	4.8	65.1%	0.7	37.1%	3.7
Ours	Point	**22.7%**	**4.3**	**23.1%**	**4.5**	**65.9%**	**0.5**	**38.5%**	**3.4**

Table 1: **Comparson on SHREC19, SURREAl, TOSCA and SMAL benchmarks.** Acc means the correspondence accuracy at an error tolerance of 0.01, while err refers to the average correspondence error. Higher accuracy and lower error reflect a better result. The datasets in brackets indicate the training sets used for the experiments, e.g., SHREC'19 (SURREAl) indicates training on the SURREAL dataset and testing on the SHREC'19 dataset.

autoencoder. The adaptive adversarial mask generator produces masks with a 40% coverage. The feature dimensions of the mask token and encoder are set to 512. The entropic regularization (ϵ) in optimal transport is configured at 10. The coefficients (λ_1, λ_2, λ_3) in Equation 10 are 1, 1, and 0.5, respectively. Our implementation utilizes CUDA 11.1 and PyTorch 1.10.1 (Paszke et al. 2017), running on a GeForce RTX 3090 device. Model training employs the AdamW (Loshchilov and Hutter 2017) optimizer with a learning rate 5e-4, a weight decay 5e-4, and a batch size of 4 for 300 epochs. More implementation details will be provided in the supplementary material.

Comparison on Human Datasets

Following previous works (Lang et al. 2021; He et al. 2023), the human datasets consist of the SURREAL and SHREC'19 datasets. There are two validation setups: Intra-dataset and cross-dataset. Specifically, Intra-dataset refers to training on the SHREC'19 training set and testing on the SHREC'19 test set. Cross-dataset involves training on the SURREAL training set and testing on the SHREC'19 test set to evaluate the model's generalization performance.

Intra-dataset evaluation. The results in Table 1 indicate that our approach achieves a 3.4% accuracy improvement and a 0.6 cm error reduction on the SHREC'19 dataset, establishing it as the new state-of-the-art method. Additional experimental results under various error tolerances can be found in Figure 1 of the supplementary materials, which demonstrate the superiority of AMIGO over other methods under the same validation setup.

Cross-dataset evaluation. Table 1 demonstrates our method achieves an accuracy improvement of 1.6% on SURREAL, with a decrease in error by 0.1 cm. This showcases excellent generalization capabilities.

Comparison on Animal Datasets

The animal datasets, specifically SMAL and TOSCA, undergo two distinct validation settings: intra-dataset and cross-dataset. Intra-dataset involves training on the TOSCA training set and testing on the TOSCA test set. Cross-dataset entails training on the SMAL training set and evaluating on the TOSCA test set to assess the model's generalization performance.

Intra-dataset evaluation. The results in Table 1 demonstrate that our approach achieves a significant accuracy improvement of 13.6% compared to the existing state-of-the-art methods on the TOSCA dataset, with a minimal error of only 0.5 centimeters. This establishes our method as the new state-of-the-art method. Additional experimental results under various error tolerances can be found in Figure 2 of the supplementary materials.

Cross-dataset evaluation. Regarding generalization performance, Table 1 indicates a 2.1% improvement in accuracy and a 0.5 reduction in error for our method compared to the current state-of-the-art method SE-ORNet (Deng et al. 2023) on the SMAL dataset.

Ablation Study

Evaluation of different model designs. In this section, we undertake comprehensive ablation studies on the SHREC'19 dataset to evaluate the efficacy of each design. Table 2 illustrates the model's performance across various designs.

DS	ASMAE	AAMG	GO	SHREC'19	
				acc ↑	err ↓
✗	✗	✗	✗	19.3%	4.9
✓	✗	✗	✗	16.5%	6.6
✓	✓	✗	✗	20.9%	4.6
✓	✓	✓	✗	21.6%	4.5
✓	✓	✗	✓	21.9%	4.4
✓	✓	✓	✓	**22.7%**	**4.3**

Table 2: **Ablation study on different designs on SHREC'19.** DS refers to downsampling. ASMAE stands for Adaptive Siamese Masked Autoencoder, AAMG refers to the Adaptive Adversarial Mask Generator, and GO represents Global Optimization.

Mask	Ratio	SHREC'19			
		acc(0.01) ↑	acc(0.05) ↑	acc(0.10) ↑	err ↓
random	20%	19.29%	46.71%	63.08%	4.68
random	40%	21.86%	50.23%	66.52%	4.43
random	60%	21.31%	48.92%	64.77%	4.49
random	80%	20.67%	47.42%	64.19%	4.55
AAMG	20%	19.85%	47.03%	63.62%	4.51
AAMG	40%	**22.72%**	**51.80%**	**69.44%**	**4.32**
AAMG	60%	22.16%	50.87%	67.64%	4.38
AAMG	80%	21.52%	49.32%	65.89%	4.43

Table 3: **Ablation studies on different masking strategies and ratios in ASMAE.** Random refers to a random mask, and AAMG denotes the adaptive adversarial mask generator. Ratio represents the ratio of mask.

Specifically, the results from the first and second rows indicate that simple downsampling can compromise the representational capability of the encoder. A comparison between the second and third rows highlights that ASMAE effectively assists the encoder in learning distinctive representations and inter-correspondence of downsampled sparse point clouds, resulting in outstanding accuracy and error improvement. The use of the adaptive adversarial mask generator to mask challenging parts and facilitate recovery in the autoencoder yields a corresponding accuracy increase of 0.7%, as indicated in the third row. The global optimization module further enhances correspondence accuracy by an additional 1.1% by planning correspondences from a global perspective. Notably, even in the absence of AAMG, global optimization proves highly advantageous.

Effectiveness of the ASMAE. In Table 3, we experimentally evaluate various mask strategies and ratios for ASMAE. Compared to MAE's commonly used random masks, AAMG achieves superior results at different ratios by adaptively selecting masked points. Different mask ratios also influence the performance of ASMAE, and ultimately, we achieve the best results using a 40% mask ratio. As depicted in Figure 5, we performed visual experiments to validate the restoration of the missing structure in the source by aggregating the complete target structure and establishing correspondences in this process. In the second column, the source point cloud constructed during the $\mathcal{L}_{mae}$ computation is presented, randomly showcasing three restored points by mask tokens color-coded in red, green, and blue. The third column exhibits the attention weights of mask tokens on the target point cloud, emphasizing the top five target points with the maximum attention weights. The visual results strongly indicate a correspondence between the restored source parts and the focused target parts by mask tokens, robustly affirming our viewpoint. Beyond qualitative results, quantitative experiments are conducted on mask token correspondence accuracy using the TOSCA dataset, as illustrated in Table 4. Specifically, we evaluated the accuracy (Acc_{mae}) of correctly establishing correspondences between the source point recovered by a mask token and the target points it attended to under different $\mathcal{L}_{mae}$ losses, including cross construction loss (CC) and mean square error loss (MSE). We apply various point thresholds ($\mathbf{T}$), considering the correspondence correct if the top $\mathbf{T}$ points in the target region most attended by the mask token include the recovered corresponding points from the source. The results further validate that our decoder not only aids in recovering source shape by the mask token but also facilitates correct correspondences with the target shape.

$\mathcal{L}_{mae}$	$\mathbf{T}$	Acc_{mae} ↑	$\mathcal{L}_{mae}$	$\mathbf{T}$	Acc_{mae} ↑
CC	1	68.3%	CC	10	**92.5%**
CC	5	86.9%	MSE	5	81.3%

Table 4: **Effectiveness of the Decoder.** AMIGO uses "CC".

Settings	SHREC'19		Settings	SHREC'19	
	acc ↑	err ↓		acc ↑	err ↓
Ind. matching	21.61%	4.53	MSE loss	22.38%	4.39
Global matching	**22.72%**	**4.32**	CE loss	**22.72%**	**4.32**

Table 5: Ablation studies on the Global Optimization.

Figure 5: **Visualization of attention on mask tokens towards the target shape.** Source and Target refer to the downsampled point cloud pair. Source(M) denotes the cross-constructed source point cloud in the ASMAE. The Attention illustrates the attention maps of three mask tokens (in red, green, and blue) on the target point cloud.

Effectiveness of the global optimization. As shown in Table 5, Global Matching with optimal transport outperforms independent matching. In order to leverage optimal transport for guiding the learning of point cloud similarity matrices, this paper employs the Cross-Entropy loss to align point cloud pairs' similarity relations with global planning outcomes. Using alternative loss functions, such as MSE loss, is also feasible. As demonstrated in the fourth row, the Cross-Entropy loss yields slightly better performance.

Conclusion

In this paper, we introduce an adaptive siamese masked autoencoder with global optimization for unsupervised point cloud shape correspondence. To our knowledge, AMIGO is the first siamese point cloud MAE framework, improving the distinctive modeling of point clouds and establishing pairwise correspondences. We present a global optimization module that utilizes optimal transport for optimization to tackle the challenge of multiple-to-one erroneous matching in shape correspondence. Extensive experiments conducted on four benchmarks and real, partial-shape datasets showcase the superior performance of AMIGO.

Acknowledgments

We thank Professor Tianzhu Zhang for his guidance and Wenfei Yang for his assistance in this research.

References

Abou Zeid, K.; Schult, J.; Hermans, A.; and Leibe, B. 2023. Point2Vec for Self-Supervised Representation Learning on Point Clouds. *arXiv e-prints*, arXiv–2303.

Borges, T. M.; Garcia, D. C.; and de Queiroz, R. L. 2022. Fractional super-resolution of voxelized point clouds. *IEEE Transactions on Image Processing*, 31: 1380–1390.

Bronstein, A. M.; Bronstein, M. M.; and Kimmel, R. 2006. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5): 1168–1172.

Bronstein, A. M.; Bronstein, M. M.; and Kimmel, R. 2008. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media.

Brown, B. J.; and Rusinkiewicz, S. 2007. Global non-rigid alignment of 3-D scans. In *ACM SIGGRAPH 2007 papers*, 21–es.

Cao, D.; and Bernard, F. 2023. Self-Supervised Learning for Multimodal Non-Rigid 3D Shape Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17735–17744.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Corman, E.; Ovsjanikov, M.; and Chambolle, A. 2014. Supervised descriptor learning for non-rigid shape matching. In *European conference on computer vision*, 283–298. Springer.

Cosmo, L.; Rodola, E.; Bronstein, M. M.; Torsello, A.; Cremers, D.; Sahillioglu, Y.; et al. 2016. SHREC'16: Partial matching of deformable shapes. *Proc. 3DOR*, 2(9): 12.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Deng, J.; Lu, J.; and Zhang, T. 2024. Unsupervised Template-assisted Point Cloud Shape Correspondence Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5250–5259.

Deng, J.; Lu, J.; and Zhang, T. 2025. Diff3DETR: Agent-Based Diffusion Model for Semi-supervised 3D Object Detection. In *European Conference on Computer Vision*, 57–73. Springer.

Deng, J.; Wang, C.; Lu, J.; He, J.; Zhang, T.; Yu, J.; and Zhang, Z. 2023. SE-ORNet: Self-Ensembling Orientation-aware Network for Unsupervised Point Cloud Shape Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5364–5373.

Deprelle, T.; Groueix, T.; Fisher, M.; Kim, V.; Russell, B.; and Aubry, M. 2019. Learning elementary structures for 3d shape generation and matching. *Advances in Neural Information Processing Systems*, 32.

Donati, N.; Sharma, A.; and Ovsjanikov, M. 2020. Deep geometric functional maps: Robust feature learning for shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8592–8601.

Dutt, N. S.; Muralikrishnan, S.; and Mitra, N. J. 2024. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4494–4504.

Ge, Z.; Liu, S.; Li, Z.; Yoshie, O.; and Sun, J. 2021. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 303–312.

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. 3d-coded: 3d correspondences by deep deformation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 230–246.

Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R. R.; and Hu, S.-M. 2021. Pct: Point cloud transformer. *Computational Visual Media*, 7(2): 187–199.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*.

He, J.; Deng, J.; Zhang, T.; Zhang, Z.; and Zhang, Y. 2023. Hierarchical Shape-Consistent Transformer for Unsupervised Point Cloud Shape Correspondence. *IEEE Transactions on Image Processing*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Huang, Q.-X.; Adams, B.; Wicke, M.; and Guibas, L. J. 2008. Non-rigid registration under isometric deformations. In *Computer Graphics Forum*, volume 27, 1449–1457. Wiley Online Library.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Lachat, E.; Landes, T.; and Grussenmeyer, P. 2019. Comparison of point cloud registration algorithms for better result assessment–towards an open-source solution. In *ISPRS TC II Mid-term Symposium "Towards Photogrammetry 2020*, volume 42, 551–558. Copernicus Publications.

Lang, I.; Ginzburg, D.; Avidan, S.; and Raviv, D. 2021. Dpc: Unsupervised deep point correspondence via cross and self construction. In *2021 International Conference on 3D Vision (3DV)*, 1442–1451. IEEE.

Li, Z.; Ai, Y.; Lu, J.; Wang, C.; Deng, J.; Chang, H.; Liang, Y.; Yang, W.; Zhang, S.; and Zhang, T. 2024. Mamba24/8D: Enhancing Global Interaction in Point Clouds via State Space Model. *arXiv preprint arXiv:2406.17442*.

Liu, W.; Sun, J.; Li, W.; Hu, T.; and Wang, P. 2019a. Deep learning on point clouds and its application: A survey. *Sensors*, 19(19): 4188.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*.

Liu, Z.; Tang, H.; Lin, Y.; and Han, S. 2019b. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, J.; Deng, J.; Wang, C.; He, J.; and Zhang, T. 2023. Query Refinement Transformer for 3D Instance Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18516–18526.

Mahmood, B.; and Han, S. 2019. 3D registration of indoor point clouds for augmented reality. In *ASCE International Conference on Computing in Civil Engineering 2019*, 1–8. American Society of Civil Engineers Reston, VA.

Marin, R.; Rakotosaona, M.-J.; Melzi, S.; and Ovsjanikov, M. 2020. Correspondence learning via linearly-invariant embedding. *Advances in Neural Information Processing Systems*, 33: 1608–1620.

Melzi, S.; Marin, R.; Rodolà, E.; Castellani, U.; Ren, J.; Poulenard, A.; Wonka, P.; and Ovsjanikov, M. 2019. Shrec 2019: Matching humans with different connectivity. In *Eurographics Workshop on 3D Object Retrieval*, volume 7, 3.

Pang, Y.; Wang, W.; Tay, F. E.; Liu, W.; Tian, Y.; and Yuan, L. 2022. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, 604–621. Springer.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Poux, F.; and Billen, R. 2019. Voxel-based 3D point cloud semantic segmentation: Unsupervised geometric and relationship featuring vs deep learning methods. *ISPRS International Journal of Geo-Information*, 8(5): 213.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Roufosse, J.-M.; Sharma, A.; and Ovsjanikov, M. 2019. Unsupervised deep learning for structured shape matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1617–1627.

Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; and Leibe, B. 2022. Mask3d for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*.

Tevs, A.; Berner, A.; Wand, M.; Ihrke, I.; and Seidel, H.-P. 2011. Intrinsic shape matching by planned landmark sampling. In *Computer graphics forum*, volume 30, 543–552. Wiley Online Library.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, C.; Deng, J.; He, J.; Zhang, T.; Zhang, Z.; and Zhang, Y. 2023. Long-short Range Adaptive Transformer with Dynamic Sampling for 3D Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5): 1–12.

Xu, Y.; Lu, Y.; and Wen, Z. 2017. Owlii Dynamic human mesh sequence dataset. In *ISO/IEC JTC1/SC29/WG11 m41658, 120th MPEG Meeting*, volume 1.

Zeng, Y.; Qian, Y.; Zhu, Z.; Hou, J.; Yuan, H.; and He, Y. 2021. CorrNet3D: Unsupervised end-to-end learning of dense correspondence for 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6052–6061.

Zhang, R.; Guo, Z.; Gao, P.; Fang, R.; Zhao, B.; Wang, D.; Qiao, Y.; and Li, H. 2022. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35: 27061–27074.

Zhang, R.; Wang, L.; Qiao, Y.; Gao, P.; and Li, H. 2023. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21769–21780.

Zhao, H.; Jiang, L.; Jia, J.; Torr, P. H.; and Koltun, V. 2021. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16259–16268.

Zuffi, S.; Kanazawa, A.; Jacobs, D. W.; and Black, M. J. 2017. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6365–6373.