DiffCorr: Conditional Diffusion Model with Reliable Pseudo-Label Guidance for Unsupervised Point Cloud Shape Correspondence

Jiacheng Deng[1], Jiahao Lu[1], Zhixin Cheng[1], Wenfei Yang[1, 2] *

[1]University of Science and Technology of China
[2]Jianghuai Advance Technology Center, Hefei, China
{dengjc,lujiahao,chengzhixin}@mail.ustc.edu.cn, yangwf@ustc.edu.cn

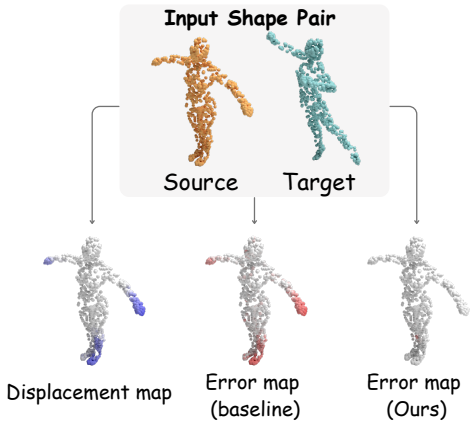
Abstract

Unsupervised point cloud shape correspondence aims to establish dense correspondences between source and target point clouds. Existing methods universally follow a one-step paradigm to obtain shape correspondence directly, but it often fails in large-scale motions of humans and animals. To address this challenge, we propose a *conditional Diffusion model with reliable pseudo-label guidance for unsupervised point cloud shape Correspondence* (DiffCorr), including a transformer-based conditional diffusion model and a reliable pseudo-label generator. The proposed DiffCorr enjoys several merits. Firstly, the transformer-based conditional diffusion model implements a coarse-to-fine optimization for coarse correspondences. Secondly, we design a reliable pseudo-label generator to provide high-quality pseudo-labels for training. Extensive experiments on four human and animal datasets demonstrate that DiffCorr surpasses state-of-the-art methods and exhibits favorable generalization capabilities.

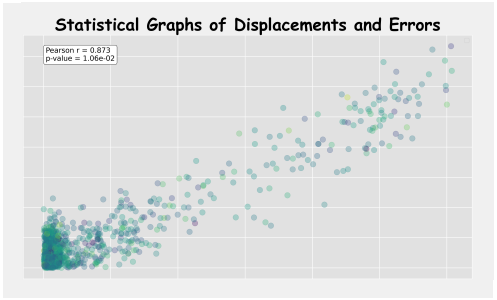
Introduction

Point cloud shape correspondence aims to identify dense mappings between two non-rigid point clouds with deformable shapes. The shape correspondence is essential for various practical applications, such as articulated motion transfer (Noguchi et al. 2022) and shape editing (Zwicker et al. 2002). However, non-rigid bodies exhibit significant motion variability. The motions lead to large-scale displacements between corresponding parts across shapes. Besides, as a primitive representation of 3D space, the sparse and irregular nature of point clouds introduces significant local noise, leading to neighborhood perturbations.

To address the aforementioned challenges, a variety of point cloud shape correspondence methods (Zeng et al. 2021; Lang et al. 2021; Deng et al. 2023; He et al. 2023) have been developed. To reduce the burden of resource-intensive labeling in fully supervised methods (Deprelle et al. 2019; Marin et al. 2020), researchers are increasingly focusing on unsupervised methods that leverage unlabeled data for model training and inference. Current unsupervised point cloud shape correspondence methods (Zeng et al. 2021; Lang et al. 2021; Deng et al. 2023; He et al. 2023) achieve state-of-the-art results by directly obtaining matching outcomes by independently calculating similarity scores between point embeddings. This one-step matching demonstrates performance improvements in areas with minor motion. However, the one-step paradigm may struggle to precisely match corresponding points with large-scale displacements, providing only an approximate estimate.



(a) Difficult to Predict Large-Scale Displacements



(b) Strong Positive Correlation between Displacement and Error

Figure 1: The correlation between displacement and error. Figure (a) visualizes the point-to-point displacement map between the source and target point clouds (the bluer, the larger) and the error maps predicted by the baseline method and our DiffCorr (the redder, the larger). Figure (b) is a scatter plot showing the strong statistical correlation between displacement and error in the point cloud.

*Corresponding Author

By examining previous point-based shape correspondence methods, we identify two key issues that should be addressed to achieve more accurate shape correspondences: 1) *How to handle large motion displacements in shape pairs?* The accuracy bottleneck in most existing methods lies in their inability to precisely predict correspondences in parts with large motion displacements, leading to substantial error accumulation. As shown in Figure 1, the visual consistency and strong positive correlation with statistical data further validate the dilemmas current methods face in large motion displacements. Therefore, we should design a coarse-to-fine correspondence prediction approach to replace the one-step paradigm, leveraging multi-step optimization to converge to accurate corresponding points gradually. 2) *How to guide the direction of multi-step optimization under unsupervised conditions?* In the absence of ground truth labels to accurately guide optimization directions, designing a selection strategy to identify high-quality landmark points as guides for multi-step optimization direction is particularly critical.

To achieve the objectives above, we propose a *conditional Diffusion model with reliable pseudo-label guidance for unsupervised point cloud shape Correspondence* (DiffCorr), which comprises a Transformer-based Conditional Diffusion Model (T-CDM), a Reliable Pseudo-Label Generator (RPLG), and a point cloud encoder within a unified architecture. The transformer-based conditional diffusion model adopts DDIM (Sundararaman, Pai, and Ovsjanikov 2022) framework with a structure-aware point cloud transformer as the backbone, facilitating multi-step denoising refinement within the local areas of initial correspondence locations. The conditions include two components: the local cost and the initial flow. The local cost is generated by inputting the K-nearest neighbor (k-NN) similarity and coordinate offsets of mapped points into a compact DGCNN (Wang et al. 2019) network to obtain local embeddings. The initial flow is the coordinate flow of initial corresponding points derived from the point similarity matrix. In the reliable pseudo-label generator, we design an effective high-quality corresponding point pair selection strategy to identify landmark points that guide the denoising direction in the diffusion model.

In summary, our contributions are as follows: (i) We introduce a novel conditional diffusion model with reliable pseudo-label guidance for unsupervised point cloud shape correspondence. (ii) The proposed transformer-based conditional diffusion model leverages conditions to overcome the bottleneck for large motion displacements. We develop a reliable pseudo-label generator that effectively filters high-quality corresponding point pairs without manual annotation. (iii) Extensive experimental results on four standard benchmarks, including SURREAL, SHREC, SMAL, and TOSCA, demonstrate that the proposed model surpasses current state-of-the-art methods.

Related Work

Learning-based point cloud representation. Learning-based point cloud representation methods are broadly divided into convolution-based (Su et al. 2015) and point-based approaches (Wu et al. 2022; Zhao et al. 2021; Wang et al. 2022; Li et al. 2024). Convolution-based works (Su et al. 2015; Chen et al. 2017) project point clouds onto 2D images to process with 2D convolutions, but this projection inevitably loses spatial information. In contrast, PointNet++ (Qi et al. 2017) advances this concept by incorporating local information. DGCNN (Wang et al. 2019) introduces EdgeConv (Wang et al. 2019) blocks, which dynamically update neighborhood information based on dynamic graphs, ultimately delivering improved performance in point cloud analysis. On the other hand, PCT (Guo et al. 2021) enhances input embedding with the support of farthest point sampling and nearest neighbor search and uses self-attention to capture global features. Our DiffCorr designs a transformer backbone tailored for the point cloud shape correspondence, aiding in learning representations and denoising local correspondences.

Shape correspondence. The shape correspondence task aims to find correspondences between two deformable shapes, which can be categorized into spectral-based and point-based methods based on the input data type. Spectral-based methods attempt to design data-agnostic shape correspondence methods to leverage the simplicity and preprocessing-free nature of point cloud data. Cao et al.(Cao and Bernard 2023) transfer rich structural information from mesh data to point clouds. Jiang et al.(Jiang, Sun, and Huang 2023) use a mesh template to predict shape correspondence between two point clouds. Unfortunately, these methods still rely on mesh data and connectivity information during training and inference. In contrast, point-based methods directly process point clouds without connectivity information. DPC (Lang et al. 2021) and SE-ORNet (Deng et al. 2023) compute point embedding similarities from the DGCNN (Wang et al. 2019) network to obtain shape correspondences directly. Diff3f (Dutt, Muralikrishnan, and Mitra 2024) distills diffusion features from foundational image models onto input shapes. Existing point-based methods struggle to handle large-scale motions effectively using this one-step correspondence strategy. Therefore, our proposed DiffCorr employs a coarse-to-fine correspondence strategy, designing a conditional diffusion model to handle large-scale motion by fine-tuning local correspondences.

Diffusion model. Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) have been extensively researched due to their powerful generative capabilities. Building on these foundations, conditional diffusion models have been developed to leverage auxiliary conditions for more controlled image synthesis. InstructPix2Pix (Brooks, Holynski, and Efros 2023) trains a conditional diffusion model using paired images and textual instructions. Moreover, the exceptional performance of diffusion models has been successfully applied to discriminative tasks (Nam et al. 2023; Ji et al. 2023) such as image segmentation (Chen et al. 2023b; Gu, Chen, and Xu 2024; Ji et al. 2023), depth estimation (Saxena et al. 2023; Kim et al. 2022; Duan, Guo, and Zhu 2023), object detection (Chen et al. 2023a; Ho et al. 2024; Deng, Lu, and Zhang 2025), and pose estimation (Tevet et al. 2022; Holmquist and Wandt 2023), demonstrating significant performance improvements. Our DiffCorr is pioneering in applying a conditional diffusion model to the unsupervised

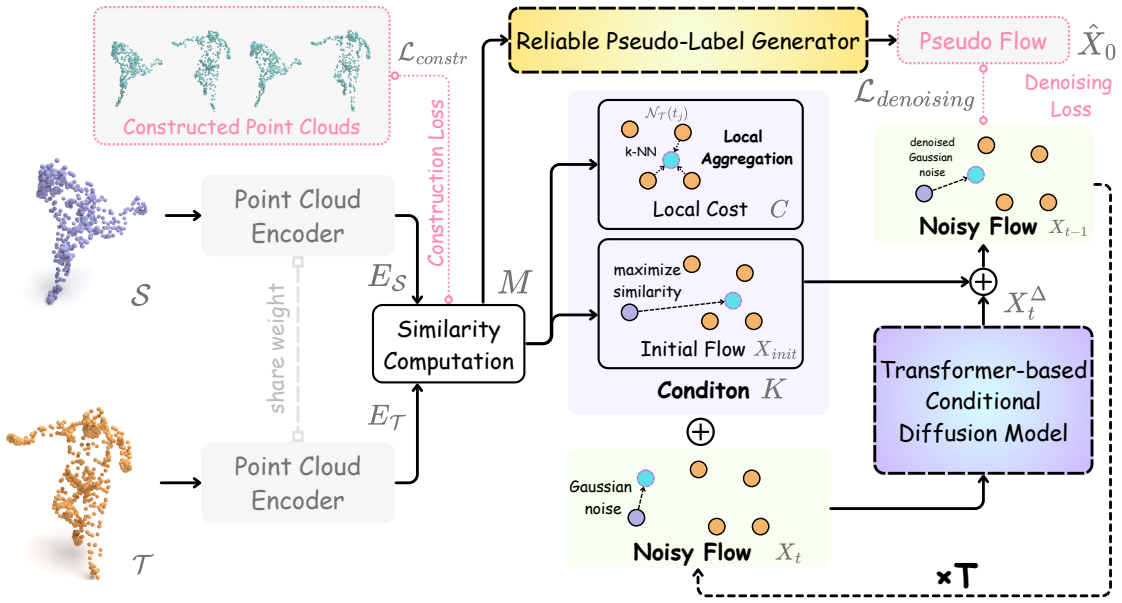


Figure 2: The framework of DiffCorr. The paired point clouds are first fed into a shared *point cloud encoder* to compute feature similarity. Based on this similarity, the initial flow and local cost are established as conditions for the *transformer-based conditional diffusion model*. The noisy flow undergoes a multi-step denoising process to obtain refined point cloud shape correspondences, supervised by the pseudo flow generated by the *reliable pseudo-label generator*. The network optimization is guided by the construction loss and denoising loss.

point cloud shape correspondence task.

Method

In this section, we first introduce the preliminaries, including the problem definition and the diffusion model. Following this, we provide an overview of our DiffCorr framework. Then, we elaborate on the details of each module and conclude with the model training and inference.

Preliminaries

Problem definition. Given two point clouds with different shapes (source point cloud $\mathcal{S} \in \mathbb{R}^{N\times 3}$ and target point cloud $\mathcal{T} \in \mathbb{R}^{N\times 3}$, where N is the number of points), our goal is to obtain a dense mapping $X^* : \mathcal{S} \to \mathcal{T}$ that matches each point $s_i \in \mathcal{S}$ of the source point cloud with its corresponding point $t_j \in \mathcal{T}$ in the target point cloud.

Conditional diffusion model. In the conditional diffusion model, the data distribution is approximated by recovering a data sample from the Gaussian noise through an iterative denoising process. Given a sample X_0, it is transformed to X_t at a time step $t \in \{T, T-1, \ldots, 1\}$ through the forward diffusion process, which consists of Gaussian transition at each time step $q(X_t|X_{t-1}) := \mathcal{N}(\sqrt{1-\beta_t}X_{t-1}, \beta_t I)$, where I is the identity matrix and β_t is the pre-defined variance at time step t. The forward diffusion process is formulated as follows:

$$X_t = \sqrt{\alpha_t}X_0 + \sqrt{1-\alpha_t}Z, \quad Z \sim \mathcal{N}(0, I), \qquad (1)$$

where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$. After training, we can sample data from the learned distribution through iterative denoising with the pre-defined range of time steps, called the reverse diffusion process, following the non-Markovian process of DDIM (Song, Meng, and Ermon 2020), which is parametrized as another Gaussian transition:

$$p_\theta(X_{t-1}|X_t) := \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \sigma_\theta(X_t, t)I). \qquad (2)$$

To this end, the diffusion network $\mathcal{F}_\theta(X_t, t; K)$ predicts the denoised sample $\hat{X}_{0,t}$ given X_t, t and condition K. This network is trained to learn the denoising function $\mathcal{F}_\theta$ by minimizing the difference between the predicted and true samples at each step. One step in the reverse diffusion process can be formulated such that

$$X_{t-1} = \sqrt{\alpha_{t-1}}\mathcal{F}_\theta(X_t, t; K) + \sqrt{\frac{1-\alpha_{t-1}-\sigma_t}{1-\alpha_t}} \cdot (X_t - \sqrt{\alpha_t}\mathcal{F}_\theta(X_t, t; K)) + \sigma_t Z, Z \sim \mathcal{N}(0, I), \qquad (3)$$

where σ_t is the covariance value of Gaussian distribution at time step t. This iterative denoising process can be viewed as finding $X^* = \arg\max_X \log p(X|K)$ through the relationship between the conditional sampling process of DDIM (Song, Meng, and Ermon 2020) and conditional score-based generative models (Batzolis et al. 2021).

Framework overview

The overall framework of DiffCorr is illustrated in Figure 2. The source point cloud $\mathcal{S}$ and target point cloud $\mathcal{T}$ are first input into a *point cloud encoder* to obtain point embeddings ($E_{\mathcal{S}}$ and $E_{\mathcal{T}}$). The cosine similarity between these point embeddings is calculated to produce a similarity matrix M. As in previous work, we employ a construction loss to aid in

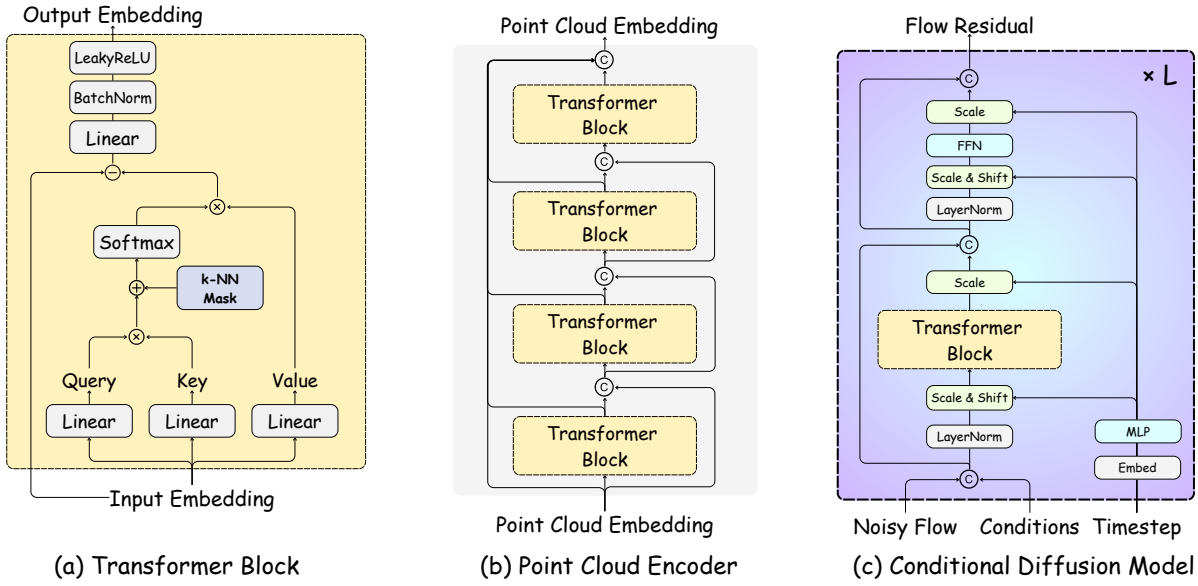


Figure 3: The architectures of transformer block, point cloud encoder and conditional diffusion model.

network training. Next, we identify the initial flow X_{init} from the source point cloud to the target point cloud by using the soft-argmax operation on similarity. We then aggregate local similarity and relative structural information from the target points around the initial corresponding points to compute the local cost C. The random Gaussian noisy flow X_t and conditions K (initial flow and local cost), are concatenated and input into the *transformer-based conditional diffusion model*. The output flow residual X_t^{Δ} is added to the initial flow to obtain the next step's noisy flow X_{t-1}. In the unsupervised setting, we design a novel *reliable pseudo-label generator* to produce pseudo flow $\hat{X}_0$ to supervise the training of the diffusion model. The details of each part will be elaborated in sequence.

Point cloud encoder

The point cloud encoder adopts a transformer architecture, leveraging the alignment between the transformer's input invariance and the unordered nature of point cloud distributions to extract structure-aware point embeddings. Specifically, as shown in Figure 3(b), our point cloud encoder architecture, similar to HSTR (He et al. 2023), comprises four transformer blocks. For the shape correspondence task, as illustrated in Figure 3(a), we restrict the receptive field of each point within the k-NN neighborhood in the attention mechanism. This strategy ensures the long-range perception capability while enhancing the modeling of local structures.

Transformer-based conditional diffusion model

Condition. The conditional diffusion model in DiffCorr uses the initial flow X_{init} and the local cost C as conditions. For any point $s_i \in \mathcal{S}$, the most similar point $t_j \in \mathcal{T}$ can be identified as the initial corresponding point using the similarity matrix M. The initial flow $x_i \in X_{init}$ measures the displacement of t_j relative to s_i in 3D space, i.e., $x_i = t_j - s_i$. The local cost $c_i \in C$ aggregates the similarity and relative displacement of the k nearest neighbors of t_j in $\mathcal{T}$. The specific calculation is as follows:

$$c_i = \max_{k \in \mathcal{N}_{\mathcal{T}}(t_j)} \mathrm{MLP}(\mathrm{Concat}(\mathrm{Concat}(t_k, m_{ik}) - \mathrm{Concat}(t_j, m_{ij}), (\mathrm{Concat}(t_j, m_{ij})))) \tag{4}$$

where $m_{ik} \in M$, and $\mathcal{N}_{\mathcal{T}}(t_j)$ represents the k-nearest neighbors of t_j in the target point cloud $\mathcal{T}$. Finally, our conditions K are composed of the concatenation of the initial flow X_{init} and the local cost C, i.e., $K = \mathrm{Concat}(X_{init}, C)$.

Noisy flow. During training, the initial noisy flow X_T is diffused according to Equation 1. While during inference, it is replaced with random Gaussian noise. At time step t, the noisy flow X_t and conditions K are input into the transformer-based conditional diffusion model to obtain the flow residual X_t^{Δ}. The initial flow X_{init} and the flow residual X_t^{Δ} are then added to produce the denoised noisy flow X_{t-1} for the next time step $t-1$, i.e., $X_{t-1} = X_{init} + X_t^{\Delta}$.

Architecture of conditional diffusion model. The detailed architecture of the proposed transformer-based conditional diffusion model is shown in Figure 3(c). The noisy flow X_t and conditions K are concatenated and input into the diffusion model. Additionally, we introduce the timestep t using the adaLN-Zero method (Peebles and Xie 2023), where scaling and shifting are applied before and after transformer block and feed-forward network (FFN). Specifically, we regress dimension-wise scale $(\alpha_1, \alpha_2, \gamma_1, \gamma_2)$ and shift (β_1, β_2) parameters from the embedding vector of t. The calculation is as follows:

$$(\alpha_1, \alpha_2, \gamma_1, \gamma_2), (\beta_1, \beta_2) = \mathrm{MLP}(\mathrm{Embed}(t)), \tag{5}$$

where MLP is a multilayer perceptron and Embed is the encoding of the timestep.

Reliable pseudo-label generator

Another major challenge in the unsupervised point cloud shape correspondence task is the lack of annotated labels. To facilitate the training of the conditional diffusion model, we design a reliable pseudo-label generator that fully exploits the information from the similarity matrix to generate reliable pseudo-labels. First, we model the uncertainty of each point through the variance values of the similarity vectors. Specifically, for each source point s_i, the total variance $\sigma^2(s_i)$ can be calculated as:

$$\sigma^2(s_i) = \mathrm{Tr}(\Sigma(m_i)), \tag{6}$$

where $\Sigma(m_i)$ is the covariance matrix of the $m_i \in M$, and $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. High variance indicates multiple or diffuse modes, signifying an unreliable prediction. This uncertainty (variance) σ^2 is used to weight the denoising loss.

Due to the presence of significant noise and many-to-one matching cases in the similarity matrix calculated directly from point embeddings, we initially aim to model the point cloud matching as an optimal transport problem to address the many-to-one mismatches. Given the target point representations $E_{\mathcal{T}} \in \mathbb{R}^{N\times C}$, we are interested in mapping the source point representations $E_{\mathcal{S}} \in \mathbb{R}^{N\times C}$ to the target points. We denote this mapping as X^{ot} with a similarity cost matrix $(1 - E_{\mathcal{T}}E_{\mathcal{S}}^{\top})$. The optimal transportation plan $X^{ot} \in \mathbb{R}^{N\times N}$ is obtained by minimizing the similarity cost with a regularization term:

$$X^{ot} = \min_{\mathbf{X}\in\mathcal{X}} \mathrm{Tr}(\mathbf{X}^{\top}(1 - \mathrm{E}_{\mathcal{T}}\mathrm{E}_{\mathcal{S}}^{\top})) - \epsilon \mathrm{H}(\mathrm{X}), \tag{7}$$

where $\mathbf{H}(\mathbf{X}) = -\sum_{ij} \mathbf{X}_{ij}\log\mathbf{X}_{\mathrm{ij}}$ is the entropy function and ϵ is the coefficient. However, we discover that the hard labels from the optimal transport solution ($\epsilon = 0$) surprisingly resulted in suboptimal or even degraded performance, likely because solving for the optimal solution amplifies the noise in the similarity matrix. Therefore, we use the Sinkhorn algorithm (Cuturi 2013) to obtain an approximate solution ($\epsilon > 0$) for optimal transport for subsequent calculations. We identify the pseudo point flow X_0^p from the source point cloud to the target point cloud by using the soft-argmax operation on the approximate solution. Finally, a cycle-consistent matching strategy is designed to filter out point pairs $\hat{X}_0$ that mutually predict each other as the best K_c correspondences, which will supervise the training of the diffusion model.

Model training and inference

In the training phase, the conditional diffusion model learns the prior knowledge of the shape correspondence with the initial flow X_{init} to give a matching hint and the local cost C to provide additional local point-wise interactions. The loss function for training diffusion model is defined as:

$$\mathcal{L}_{denoising} = \mathbb{E}_{\hat{X}_0,t,Z\sim\mathcal{N}(0,I)}\left[\frac{1}{\sigma^2}\left\|\hat{X}_0 - \mathcal{F}_\theta(X_t, t; K)\right\|^2\right]. \tag{8}$$

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 two key components: the denoising loss ($\mathcal{L}_{denoising}$) and the construction loss for point cloud pairs ($\mathcal{L}_{constr}$), formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{denoising} + \lambda \cdot \mathcal{L}_{constr}, \tag{9}$$

where λ is the hyperparameter to balance the contribution of different loss terms.

During the inference phase, a Gaussian noise X_T is gradually denoised into a more accurate matching field X_0 under the given conditions through the diffusion reverse process. To account for the stochastic nature of diffusion-based models, we propose utilizing multiple hypotheses by computing the mean of the estimated multiple matching fields from multiple initializations F_T, which helps to reduce stochasticity of model while improving the shape correspondence performance.

Experiment

Experimental settings

Datasets. Experiments are mainly conducted on four datasets including SURREAL (Groueix et al. 2018), SHREC'19 (Melzi et al. 2019), SMAL (Zuffi et al. 2017) and TOSCA (Bronstein, Bronstein, and Kimmel 2008). SURREAL is a human dataset containing 230,000 training shapes. For a fair comparison, we follow previous methods (Lang et al. 2021; He et al. 2023; Deng et al. 2023) and randomly sample 2,000 shapes for the training dataset. The SHREC'19 dataset includes 44 human models, which pair to form 430 shape pairs. SMAL is an animal dataset composed of a parametric model and contains 10,000 shapes. TOSCA is another animal dataset with 80 shapes. The detailed dataset statistics are summarized in Table 1.

Dataset	SURREAL	SHREC'19	SMAL	TOSCA
# of Points	1024	1024	1024	1024
Category	Human	Human	Animal	Animal
Non-Rigidity	✓	✓	✓	✓
# of Shapes	230K	44	10K	80
Model Type	Synthetic 3D Mesh	Synthetic 3D Mesh	Parametric Model	Synthetic 3D Mesh

Table 1: Dataset information.

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 $(\mathcal{S}, \mathcal{T})$, is formulated as:

$$err = \frac{1}{N}\sum_{s_i\in\mathcal{S}} \left\|X^*(s_i) - t_{\mathbf{gt}}\right\|_2, \tag{10}$$

where $t_{\mathbf{gt}} \in \mathcal{T}$ is the ground-truth matching point to s_i. The correspondence accuracy is defined as:

$$acc(\epsilon) = \frac{1}{N}\sum_{s_i\in\mathcal{S}} \mathbb{1}\left(\left\|X^*(s_i) - t_{\mathbf{gt}}\right\|_2 < \epsilon d\right), \tag{11}$$

Method	SURREAL/SHREC'19		SHREC'19/SHREC'19		SMAL/TOSCA		TOSCA/TOSCA	
	acc↑	err↓	acc↑	err↓	acc↑	err↓	acc↑	err↓
Diff-FMaps	4.0%	7.1	-	-	-	-	-	-
3D-CODED	2.1%	8.1	-	-	0.5%	19.2	-	-
Elementary	2.3%	7.6	-	-	0.5%	13.7	-	-
CorrNet3D	6.0%	6.9	0.4%	33.8	5.3%	9.8	0.3%	32.7
DPC	17.7%	6.1	15.3%	5.6	33.2%	5.8	34.7%	2.8
HSTR	19.4%	5.6	19.3%	4.9	33.9%	5.6	52.3%	1.2
SE-ORNet	<u>21.5%</u>	<u>4.6</u>	17.5%	5.1	36.4%	3.9	38.2%	2.7
TANet	20.6%	4.8	<u>21.5%</u>	<u>4.5</u>	<u>37.1%</u>	<u>3.7</u>	<u>65.1%</u>	<u>0.7</u>
DiffCorr(ours)	**22.53±0.27%**	**4.31±0.08**	**22.16±0.35%**	**4.36±0.07**	**41.64 ± 0.56%**	**2.31± 0.12**	**66.71±1.79%**	**0.65 ± 0.03**
Improvement Δ	1.03%↑	0.29↓	0.66%↑	0.14↓	4.54%↑	1.39↓	1.61%↑	0.05↓

Table 2: Main results on four benchmarks. We compare our method with state-of-the-art works on the SURREAL, SHREC'19, SMAL, and TOSCA benchmarks at an error tolerance of 0.01. Higher accuracy and lower error indicate better performance. The datasets are divided with the training set listed before the slash and the test set after the slash. The best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

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

Implementation details. The point cloud network consists of four transformer blocks with embedding dimensions of (96, 192, 384, 768), and finally, an MLP generates 512-dimensional point embeddings. In the transformer block, the number of nearest neighbors for the k-NN mask is set to $k = 27$, while for the local cost, $k = 16$. The channel dimension of the conditional diffusion model is 512, and the number of blocks L is set to 3. For diffusion reverse sampling, we employ the DDIM sampler (Song, Meng, and Ermon 2020) and set the diffusion timestep T to 5 during both the training and sampling phases. The default number of samples for multiple hypotheses is set to 3 for evaluations on SHREC'19 and 4 for TOSCA. The entropy regularization coefficient ϵ in the Sinkhorn algorithm is set to $\frac{1}{10}$. In the cycle-consistent matching strategy, K_c is set to 12. The λ in Equation 9 is set to 0.8. Our implementation utilizes CUDA 11.3 and PyTorch 1.12.1 (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 of 5×10^{-4}, a weight decay of 5×10^{-4}, and a batch size of 4 for 300 epochs. More details will be provided in the supplementary materials.

Comparison with state-of-the-art methods

In this subsection, we compare DiffCorr with current state-of-the-art methods on human datasets (SURREAL, SHREC'19) and animal datasets (SMAL, TOSCA). Table 2 presents the experimental results on the SHREC'19 and TOSCA datasets, as well as cross-dataset generalization experiments with SURREAL training and SHREC'19 testing (SURREAL/SHREC'19) and SMAL training and TOSCA testing (SMAL/TOSCA). With few bells and whistles, the proposed DiffCorr achieves state-of-the-art performance on both human and animal benchmarks.

Experiments on human datasets. As shown in Table 2 (SHREC'19/SHREC'19), our DiffCorr achieves new state-of-the-art results, surpassing HSTR (He et al. 2023) by 2.86% in accuracy and reducing the average correspondence error by 0.54. To further validate DiffCorr's generalization capability on human datasets, we train DiffCorr on SURREAL and test it on SHREC'19. The results in Table 2 (SURREAL/SHREC'19) indicate that DiffCorr outperforms SE-ORNet (Deng et al. 2023), improving accuracy by 1.03% and reducing error by 0.29. The improvements substantially verify the generalization capability of our proposed model. In addition to these quantitative results, Figure 4 presents a qualitative comparison with other methods. The visualization results more intuitively demonstrate the superiority of our method.

Experiments on animal datasets. The results in Table 2 (TOSCA/TOSCA) demonstrate that DiffCorr also achieves new state-of-the-art results on animal datasets. DiffCorr surpasses HSTR by 14.41% in accuracy and reduces the error by 0.55 to a satisfactory 0.65 cm. To further validate DiffCorr's generalization capability on animal datasets, we train DiffCorr on SMAL and test it on TOSCA. As shown in Table 2 (SMAL/TOSCA), DiffCorr significantly outperforms the state-of-the-art method SE-ORNet, with a 5.24% increase in accuracy and a 1.59 reduction in error, robustly validating its superior generalization capability. Additionally, Figure 4 presents a qualitative comparison with other methods, vividly demonstrating the superiority of DiffCorr.

Ablation study

To gain deeper insights into the proposed method, we conduct detailed ablation studies in Table 3 to evaluate the effectiveness of each component in DiffCorr and discuss the advantages and disadvantages of different designs.

Effectiveness of the transformer-based conditional diffusion model. In Table 3, we validate the effectiveness of the transformer-based conditional diffusion model and different conditions (local cost and initial flow). Specifically, comparisons [B] vs [A], [C] vs [A] and [D] vs [A] demonstrate that our proposed diffusion model significantly improves the accuracy. After further adding conditions ([E], [F], [G]), the accuracy keeps increasing from 62.79% to 66.71%, validating that the combination of coarse correspondence prediction and local information around the target point can mutually enhance and better guide the diffusion model to obtain

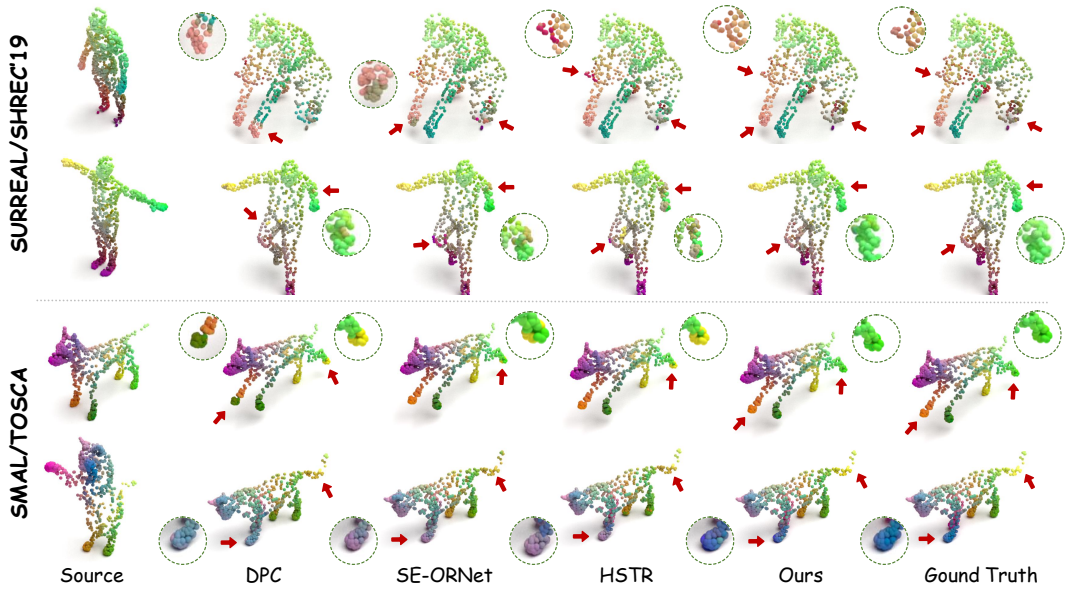


Figure 4: Visual comparison on SHREC'19 (trained on SURREAL) and TOSCA (trained on SMAL).

more robust denoising capabilities.

	T-CDM	LC	IF	OT	CC	TOSCA/TOSCA acc ↑	TOSCA/TOSCA err ↓
[A]	✗	✗	✗	✗	✗	54.58%	1.21
[B]	✓	✗	✗	✓	✗	60.82%	0.98
[C]	✓	✗	✗	✗	✓	60.15%	1.05
[D]	✓	✗	✗	✓	✓	62.79%	0.82
[E]	✓	✓	✗	✓	✓	63.65%	0.77
[F]	✓	✗	✓	✓	✓	65.31%	0.71
[G]	✓	✓	✓	✓	✓	**66.71%**	**0.65**

Table 3: Evaluation of the model with different designs on the TOSCA dataset. T-CDM denotes the transformer-based conditional diffusion model, LC stands for the local cost, IF represents initial flow, and OT and CC denote the optimal transport and cross-consistent matching strategy in the reliable pseudo-label generator, respectively.

Effectiveness of the reliable pseudo-label generator. A major challenge in leveraging the multi-step denoising optimization capability of the diffusion model for the unsupervised point cloud shape task is how to effectively constrain the learning of the diffusion model. The comparisons in Table 3 ([B] vs [C] vs [D]) fully validate the rationality of our designed optimal transport and cross-consistent matching strategy, which can effectively enhance and filter out reliable pseudo flows to constrain the learning of the conditional diffusion model.

Conclusion

In this paper, we propose a conditional diffusion model with reliable pseudo-label guidance for unsupervised point cloud shape correspondence, including a transformer-based conditional diffusion model and a reliable pseudo-label generator. Specifically, the transformer-based conditional diffusion model is introduced to refine the initial coarse predictions via denoising capabilities with initial correspondences and local structural information as conditions. The reliable pseudo-label generator filters reliable pseudo-labels based on point cloud feature similarities for training the conditional diffusion model. Extensive experiments on the SHREC'19 and TOSCA benchmarks demonstrate the superiority of DiffCorr. Cross-dataset experiments on SURREAL and SMAL further showcase its outstanding generalization capabilities.

Limitations. DiffCorr tactfully incorporates the conditional diffusion model into the unsupervised point cloud shape correspondence task and achieves new state-of-the-art results across various datasets. However, DiffCorr also faces limitations due to the slow inference speed and high randomness inherent to diffusion models. In real scanning, background elements increase our failure cases. To address these issues, we adopt a skip-step sampling strategy (Sundararaman, Pai, and Ovsjanikov 2022) that completes the denoising process in just five steps, significantly accelerating the inference speed. Additionally, we implement a multi-sample averaging strategy to mitigate fluctuations caused by randomness effectively.

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (NO.62306294 and NO.12150007), Dreams Foundation of Jianghuai Advance Technology Center (NO.2023-ZM01Z019) and Youth Innovation Promotion Association.

References

Batzolis, G.; Stanczuk, J.; Schönlieb, C.-B.; and Etmann, C. 2021. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*.

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

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

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.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023a. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19830–19843.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2023b. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 909–919.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1907–1915.

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

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.

Duan, Y.; Guo, X.; and Zhu, Z. 2023. Diffusiondepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*.

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.

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.

Gu, Z.; Chen, H.; and Xu, Z. 2024. Diffusioninst: Diffusion model for instance segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2730–2734. IEEE.

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.

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*.

Ho, C.-J.; Tai, C.-H.; Lin, Y.-Y.; Yang, M.-H.; and Tsai, Y.-H. 2024. Diffusion-SS3D: Diffusion Model for Semi-supervised 3D Object Detection. *Advances in Neural Information Processing Systems*, 36.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Holmquist, K.; and Wandt, B. 2023. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15977–15987.

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; and Luo, P. 2023. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21741–21752.

Jiang, P.; Sun, M.; and Huang, R. 2023. Non-Rigid Shape Registration via Deep Functional Maps Prior. *arXiv preprint arXiv:2311.04494*.

Kim, G.; Jang, W.; Lee, G.; Hong, S.; Seo, J.; and Kim, S. 2022. Dag: Depth-aware guidance with denoising diffusion probabilistic models. *arXiv preprint arXiv:2212.08861*.

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*.

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

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.

Nam, J.; Lee, G.; Kim, S.; Kim, H.; Cho, H.; Kim, S.; and Kim, S. 2023. Diffmatch: Diffusion model for dense matching. *arXiv preprint arXiv:2305.19094*.

Noguchi, A.; Iqbal, U.; Tremblay, J.; Harada, T.; and Gallo, O. 2022. Watch it move: Unsupervised discovery of 3d joints for re-posing of articulated objects. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3677–3687.

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.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. 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.

Saxena, S.; Kar, A.; Norouzi, M.; and Fleet, D. J. 2023. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, 945–953.

Sundararaman, R.; Pai, G.; and Ovsjanikov, M. 2022. Implicit field supervision for robust non-rigid shape matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, 344–362. Springer.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3): 415–424.

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.

Wu, X.; Lao, Y.; Jiang, L.; Liu, X.; and Zhao, H. 2022. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35: 33330–33342.

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.

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.

Zwicker, M.; Pauly, M.; Knoll, O.; and Gross, M. 2002. Pointshop 3D: An interactive system for point-based surface editing. *ACM Transactions on Graphics (TOG)*, 21(3): 322–329.