Canonical Pose Reconstruction from Single Depth Image for 3D Non-rigid Pose Recovery on Limited Datasets

Fahd Alhamazani, Paul L. Rosin, Yu-Kun Lai

[a]*Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia*
[b]*Cardiff University, Abacws Building, Cardiff, CF24 4AG, UK*
[c]*Cardiff University, Abacws Building, Cardiff, CF24 4AG, UK*

Abstract

3D reconstruction from 2D inputs, especially for non-rigid objects like humans, presents unique challenges due to the significant range of possible deformations. Traditional methods often struggle with non-rigid shapes, which require extensive training data to cover the entire deformation space. This study addresses these limitations by proposing a canonical pose reconstruction model that transforms single-view depth images of deformable shapes into a canonical form. This alignment facilitates shape reconstruction by enabling the application of rigid object reconstruction techniques, and supports recovering the input pose in voxel representation as part of the reconstruction task, utilizing both the original and deformed depth images. Notably, our model achieves effective results with using a small dataset with 300 samples in total, containing variations in shape (obese, slim and fit bodies) and gender (female and male) and size (child and adult). Experimental results on animal and human datasets demonstrate that our model outperforms other state-of-the-art methods.

Keywords: Non-rigid deformation, Canonical pose, Depth images

1. Introduction

3D reconstruction aims to turn 2D inputs such as images into 3D shapes. Most 3D reconstruction methods are designed for rigid shapes. But for non-

Email addresses: `fahd.alhamzani@nbu.edu.sa` (Fahd Alhamazani), `rosinpl@cardiff.ac.uk` (Paul L. Rosin), `laiy4@cardiff.ac.uk` (Yu-Kun Lai)

rigid objects that can bend or twist, such as humans or animals, it gets tricky. These objects can have a large range of deformation, making them hard to handle, especially for reconstruction tasks, since many training examples are required to cover the deformation space. To make the problem more manageable, an effective approach is to bring non-rigid shapes back to a default or standardised pose. This default pose is called the canonical form. Using this form can help simplify and improve various geometric processing tasks, from shape retrieval to shape reconstruction.

Canonical form refers to a normalised representation of a deformable shape such that various instances of similar objects are represented in a unified pose which removes the non-rigid deformation. This uniform representation aids in reducing variability [1], ensuring consistency, and simplifying subsequent computational processes [2]. The canonical form is commonly used in retrieval tasks, enabling us to search for and identify similar 3D models regardless of their deformations. However, current canonical form methods often prioritise discriminating between shapes but fail to retain good quality of shape appearance. These approaches typically rely on either Euclidean distance [3, 4] or geodesic distance [5] which can distort the deformed shapes. Alternatively, some works suggest other approaches like mapping the deformed shape to a template to preserve shape appearance [4]. However, these methods all assume that the input is a complete deformed shape, so cannot be applied to cases with depth image input.

In terms of non-rigid shape completion, unlike existing methods [6, 7, 8] that rely heavily on large-scale datasets to achieve accurate 3D reconstruction, our model demonstrates effective non-rigid shape reconstruction from a single depth image, utilising a considerably smaller dataset. This approach addresses the challenge of data efficiency in 3D reconstruction and highlights the model's capability to generalise accurately even with limited training samples.

In this study, we address the problem of transforming a deformable shape, represented as a single-view depth image, into its canonical form as the **first stage** of our approach. This initial step is particularly challenging, as the input does not represent a complete shape. In the **second stage**, we utilise the reconstructed canonical pose to estimate the full 3D shape of the non-rigid object. Our two-stage design makes this task more manageable, especially with limited training data, by breaking the problem down into simpler subparts. Moreover, as we will later demonstrate, the first-stage results are already sufficient for many applications such as retrieval of deformable shapes.

To address the first stage challenge, we introduce a learning-based model that converts a single depth image to a default pose. Given a 2D depth image and its corresponding mask, our model aims to produce a depth image that corresponds to the input shape in a canonical pose. Figure 1 displays an overview of the model for the first stage (canonical pose estimation), which begins with an encoder-decoder that produces high-dimensional local features. Additionally, we introduce parallel encoders utilising sparse convolution to detect neighbour sizes, thereby fusing multiscale features that contribute to preserving shape appearance. These fused features serve as a basis to generate high-dimensional attributes. Ultimately, we use an encoder-decoder model to reconstruct the canonical pose depth image.

In Stage Two, our model incorporates both a pose encoder and a shape encoder to estimate the 3D shape from the canonical pose depth image and the original input depth image. The pose encoder processes the original input depth image, while the shape encoder processes the canonical depth image to capture structural details. Finally, we employ a discriminator, where (following [9]) GAN divergence is used to smooth the volume surface and refine 3D reconstruction quality. This setting allows the model to recover the complete 3D shape with high fidelity.

Our contributions are:

- We propose a canonical pose reconstruction model, which is an end-to-end 2D network designed for the canonical pose reconstruction of single-view depth images. It comprises three components, Local Features Extractor (LF), Multi-Scale Features Extractor (MSF) and canonical pose reconstruction.

- We propose parallel encoders and a single decoder block that extract features at different scales and use a fusing decoder to decode multi-scale, high-dimensional features.

- We propose a model that estimates the full 3D shape with its recovered pose.

- The extensive experimental results on TOSCA [10] and human [11] datasets demonstrate that our model outperforms the existing state-of-the-art methods. Moreover, our model is also capable of preserving high-quality shape details while deforming shapes across different types of forms, such as humans and animals with limited datasets.

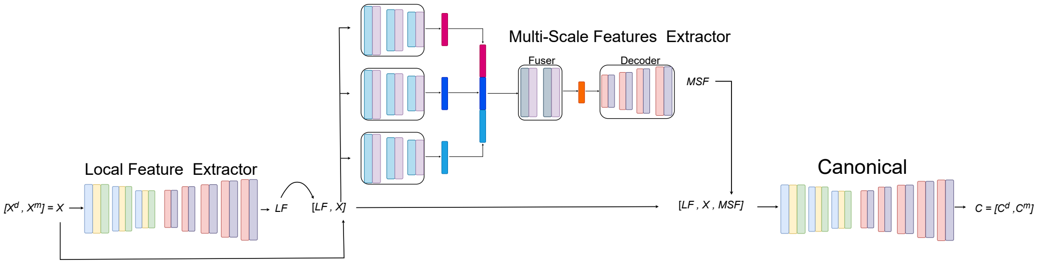

Figure 1: Overview of our first-stage model, which takes a depth image in any pose as input and outputs the canonical front-facing depth image. The Local Feature Extractor (LFE) processes the input to generate a local feature map. Then, the Multi-Scale Feature Extractor (MSFE) processes both the original input (X) and the local feature map (LF) to extract a multi-scale feature map (MSF). Finally, the Canonical Component processes all the previous maps (LF and MSF) along with the original input (X) to estimate the canonical depth image. Each component is further illustrated in the subsequent figures.

2. Related Work

As mentioned before, our method involves two stages, namely canonicalisation of deformable shapes and single-view 3D shape completion from the canonical form. Related work in these two areas is described below.

2.1. Canonicalisation of Deformable Shapes from Single Depth Input

Many tasks including 3D reconstruction and shape retrieval benefit from putting deformable shapes into some standardised poses (such as T-pose for human bodies), which are referred to as canonical forms. For example, shape retrieval is an important task that aims to find similar shapes to the query. Many methods work well on rigid bodies where all shapes have fixed pose. However, these methods may work poorly on non-rigid shapes, where the same shape can have different poses. Without a standardised pose (canonical form), determining correspondence between points on two non-rigid shapes can be ambiguous, as the geometric distances caused by pose difference are often much larger than those of different instances. Also, machine learning algorithms, especially those based on deep learning, require consistent data representation for effective training, such as learning-based 3D reconstruction. Different poses can be seen as “noise” or “variations” that can affect the learning process if not standardised through canonical forms. To solve that, a canonical form standardises the shapes to a fixed pose. In this section, we will review canonical form for non-rigid shapes techniques.

Canonical form reconstruction for non-rigid shapes has been a hot area of research due to its utility in tasks like shape matching and retrieval. Early approaches often relied on geometrically motivated transformations. Lian et al. [4] introduced a feature-preserving canonical form using Multidimensional Scaling (MDS) to transform non-rigid 3D watertight meshes. Their method segments objects into near-rigid parts and optimises alignment, preserving key features by minimising non-linear deformations. Although effective, this method can be computationally intensive, limiting its scalability to larger datasets. In contrast, Pickup et al. [3] achieved computational efficiency by using Euclidean distances between selected vertices to approximate global geodesic distances, offering a faster alternative suitable for high-resolution meshes. However, this approach is limited in adaptability, as it depends on vertex conformal factors to identify structural features, which can be insufficient for highly deformable shapes. Other efforts, like the work of Lian et al. [12], approached shape matching with image-based methods, using MDS and PCA to capture the canonical pose of objects in multi-view depth images. Though computationally lighter, these multi-view representations require extensive feature extraction, adding complexity to real-time applications.

Another major stream of works have focused on embedding techniques and multi-feature fusion to handle more complex, non-rigid deformations. Wang and Zha [13] introduced the contour canonical form, leveraging geodesic constraints to ensure isometry invariance but at a cost of increased geodesic calculations. Building on the need for more flexible approaches, Zeng et al. [14] combined canonical forms with multi-view convolutional neural networks, enhancing feature extraction through multi-feature fusion methods, though heavily reliant on extensive data. Meanwhile, Jribi and Ghorbel [15] proposed a method using geodesic distances to reference points, addressing the challenge of inelastic deformations. More recent works by [2, 16] used random walks and local commute time distance, allowing shape retrieval by segmenting objects into localised regions, which are then merged for pose-invariant canonical forms. Although these local methods improve retrieval accuracy by preserving salient features, they can struggle to maintain global shape coherence, underscoring the need for approaches that balance global structure with local detail retention.

2.2. Single-View 3D Shape Completion from Canonical Form

While many 3D shape completion methods have shown strong performance on rigid objects, they often struggle with non-rigid bodies due to the inherent variability in pose and articulation. Rigid objects typically maintain consistent structure across instances, allowing models to learn reliable geometric priors. In contrast, non-rigid shapes – such as human bodies, animals, etc. – can appear in a large range of poses, making it difficult for standard completion models to generalise effectively. Without a standardised pose or canonical form, completing the 3D shape from partial observations becomes ambiguous, as differences in pose can overshadow structural similarity. To address this challenge, recent works have explored pose normalisation or canonicalisation as a preprocessing step, enabling the model to focus on shape completion from a consistent reference frame. In our approach, we leverage a canonical depth representation from stage one as the input for 3D reconstruction, allowing the network to disentangle pose variation from geometric completion and thereby improving the accuracy and robustness of shape recovery for non-rigid objects.

Recent research has extensively explored 3D human shape reconstruction from monocular RGB or depth inputs using parametric models such as SMPL [17, 18]. Many of these approaches, while achieving impressive results for human avatars, remain limited in scope due to their reliance on human-specific priors, fixed topology, and large annotated datasets. For example, Wang et al. [17] proposed a hybrid method that fits the SMPL model to depth maps using deep dense correspondences, achieving accurate results for clothed human models by leveraging a double U-Net architecture followed by optimisation. However, their method is constrained to human anatomy and heavily depends on the SMPL mesh structure, making generalisation to non-rigid non-human shapes – such as animals – unfeasible.

Dong et al. [18] introduced PINA, an implicit neural avatar framework that reconstructs personalised human avatars from a single RGB-D sequence. Their method fuses partial observations into a canonical signed distance field and learns skinning weights and deformations jointly via global optimisation. While powerful, PINA also relies on pose-conditioned deformation fields aligned with SMPL priors and is not easily transferable to subjects with different body plans or articulation models, such as quadrupeds.

Similarly, PSHuman [19] adopts a multiview diffusion-based framework to generate textured 3D meshes from a single RGB image. It leverages SMPL-X priors for geometry stabilisation and combines cross-scale generation with explicit remeshing for high fidelity. While it achieves high-quality reconstructions and appearance consistency, it is fundamentally built upon human body assumptions, including face-body scale separation and SMPL-X initialisation, which make it unsuitable for generalised shape completion tasks in animals or unfamiliar topologies.

Xue et al. [20] proposed Neural Surface Fields (NSF), which generate animatable human models from monocular depth by learning deformation fields over a canonical implicit base shape. NSF excels in mesh coherence and efficiency, and supports arbitrary resolution outputs without retraining. Yet, like prior works, it depends on SMPL-based canonicalization and human-specific priors. Although powerful in generalising across human poses and clothing, NSF's applicability to non-human categories remains unexplored.

In contrast, our approach removes this dependency on SMPL or any human parametric model. We reconstruct a full 3D shape from a single canonical depth map without requiring large datasets, registered meshes, or human-specific priors. Moreover, our model is explicitly designed to generalise to non-human, highly deformable categories – such as animals – by learning pose-independent shape priors directly from canonicalised depth representations. This generalisation beyond the human domain makes our approach more flexible and applicable to diverse shape completion tasks where parametric models like SMPL do not exist.

3. Methodology

The canonical form involves addressing deformation by eliminating it, aiming to transform the input depth image to align with a standardized canonical form. This model comprises four main components. First, the initial component interprets the depth image to extract high-dimensional local features, which are then integrated with the original depth information (Section 3.1). Next, the model employs parallel encoders alongside a fusing decoder to generate multi-scale features (Section 3.2). These multi-scale features are subsequently concatenated with the local features, creating skip connections for the canonical pose reconstruction component and grouping local features with multi-scale features to assist in reconstructing the depth image into its canonical form (Section 3.3). Finally, we combine the original depth image with the estimated canonical depth image to reconstruct the 3D shape with pose recovery (Section 3.4).

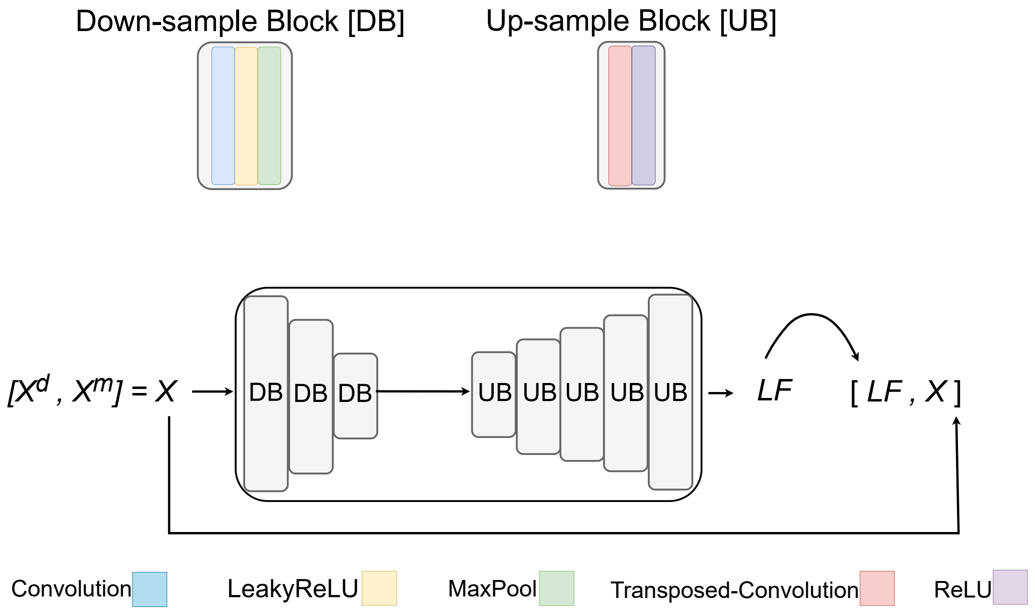

Figure 2: The Local Feature Extractor (LFE) takes the single-view depth image X^d and the corresponding mask X^m as input and produces a local feature output, of the same input size, denoted as LF

3.1. Local Feature Extractor

Given an input depth image X^d and mask X^m, where $X^d = \{x_i^d \in \mathbb{R}^{500\times500}\}$, $X^m = \{x_i^m \in \{0,1\}^{500\times500}\}$ in our implementation, the LFE component processes both in order to generate local features. The component consists of N down-sample blocks and K up-sample blocks, where $N = 3$ and $K = 5$. For the down-sample blocks, each block consists of a convolution with a kernel size of 5×5 and strides of 1×1. We use `LeakyReLU` as the activation function, and a `MaxPool` layer is employed for spatial reduction. For the up-sample blocks, the transpose-convolutions use three different kernel sizes: $[5, 3, 2]$. The output features, denoted as LF in Eq. 1, are concatenated with the original input $X^{d,m}$ as an extra channel. The network is shown in Figure 2.

$$LF = LFE(X^d, X^m) \tag{1}$$

LFE attaches local features to the original depth image and its mask, so each pixel is associated with both a local feature and a mask value. Consequently, in Section 3.3, the reconstruction component has access to both the local features and the original input depth image.

3.2. Multi-Scale Feature Extractor

The Multi-scale Feature Extractor (MSFE) described in Eq. 2 comprises three parallel encoders $E_{dilation1}$, $E_{dilation2}$ and $E_{dilation3}$.

$$MSF = MSFE(X^d, X^m, LF) \tag{2}$$

$$z_1 = E_{dilation1}(X^d, X^m, LF) \tag{3}$$

$$z_2 = E_{dilation2}(X^d, X^m, LF) \tag{4}$$

$$z_3 = E_{dilation3}(X^d, X^m, LF) \tag{5}$$

Each encoder captures a different spatial neighbourhood size owing to the inherent nature of convolutions with distinct dilation values. Specifically, the three encoders possess dilation values of 1, 2, and 3 (Eqs. 3, 4 and 5) in their convolution layers. We could not add more than 3 encoders as computation consumption exceeds GPU limits, also 1,2,3 variation is a natural way to expand. Every encoder outputs a latent code of size 1600 ($dim(z_1) = 1600$, $dim(z_2) = 1600$ and $dim(z_3) = 1600$, where $dim(\cdot)$ is the dimension of the latent code). We found that using less than 1600 for the latent code degrades the reconstruction results. When concatenated, this results in a latent code with a length of 4800.

These parallel encoders handle pixels from different scales, thereby yielding multi-scale features. To fuse these multi-scale features, we use a single decoder, as described in Eq. 6.

$$MSF = D_{fuser}(z_1, z_2, z_3) \tag{6}$$

As an initial step, the 4800D latent codes are processed through two MLP layers to identify inter-code relationships, ultimately generating 500D latent codes. Subsequently, six transpose-convolutions are applied. Following each convolution, a `ReLU` activation function is employed. The output MSF has the spatial resolution aligned with the original input size. This design enables the association of multi-scale features with each input pixel. The overview of *MSF* is shown in figure 3.

3.3. Canonical Pose Depth Reconstruction

Deformation involves transforming a shape from any pose to a default pose. In terms of an image, this means shifting the pixels to recreate a canonical pose. However, conventional convolution cannot adequately attend

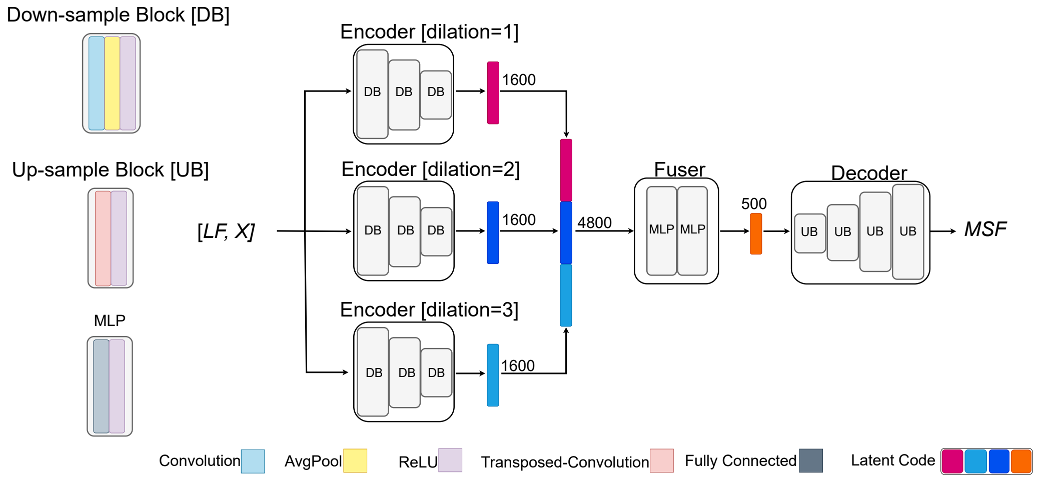

Figure 3: The model takes as input the original depth X^d, its mask X^m where $[X^d, X^m] = X$, and the local feature output Y_{LFE}. It features three encoders, each having a distinct dilation rate, with each encoder made up of down-sample blocks. Following the encoders, the latent codes are concatenated and passed through a fuser for inter-mapping. The subsequent decoder consists of up-sample blocks, culminating in the reconstructed multi-scale features, denoted as MSF.

to long dependencies. As a solution, we generate both local and multi-scale features of the same size as the input image, allowing the reconstruction component in Eq. 7 to access both feature types for each pixel.

Similar to the LF component, the reconstruction component incorporates four channels: the original input and its mask, local feature data generated by the LF component, and multi-scale features produced by MSF. Note that combining features from different stages of the model helps reduce vanishing gradients. The reconstruction component comprises N down-sample blocks and K up-sample blocks, where $N = 3$ and $K = 5$. Each down-sample block consists of a convolution layer, followed by a `LeakyReLU` and a `MaxPool` layer, with kernels of size 5 and stride 1. On the other hand, each up-sample block features a transpose-convolution and a `ReLU` layer. The final output from the canonical pose component $C^{d,m}$ is a reconstructed depth image alongside a reconstructed mask, where $C^d = \{c_i^d \in \mathbb{R}^{500\times500}\}, \quad C^m = \{c_i^m \in \{0,1\}^{500\times500}\}$ in our implementation. The overview of the reconstruction component is shown in figure 4.

$$C^{d,m} = canonical(X^d, X^m, LF, MSF) \tag{7}$$

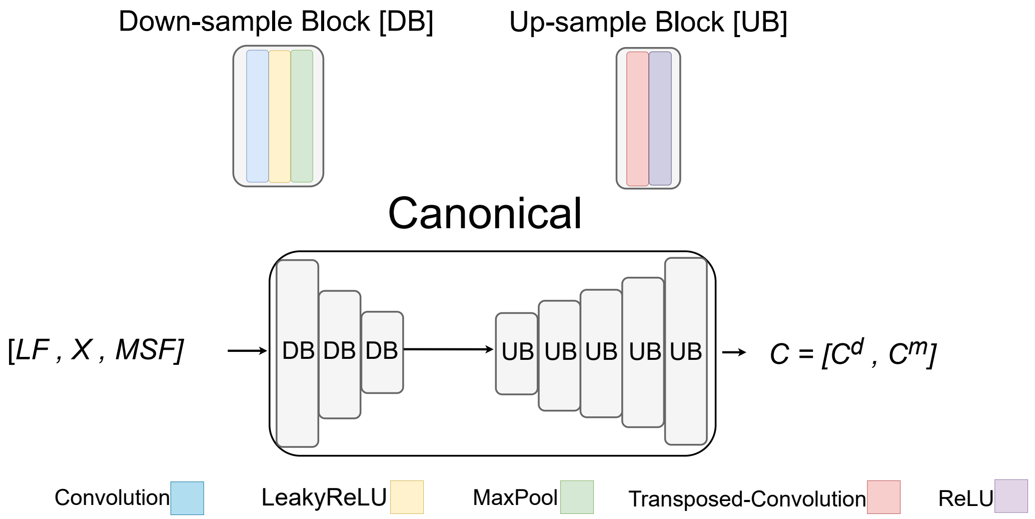

Figure 4: The canonical reconstruction component leverages the original input X, the LFE output LF, and the MSFE output MSF. The model uses these inputs to determine the canonical form C which consists of canonical form depth image C^d and its mask C^m.

3.4. Pose Recovery

After reconstructing the default pose, the next stage focuses on 3D volume shape reconstruction, using both the original pose depth image $X^{d,m}$ and the reconstructed default pose $C^{d,m}$. The architecture consists of two encoders and a decoder, collectively forming the generator in our GAN framework, as shown in Figure 5. Specifically, the pose encoder Eq. 9 processes the original input depth image $X^{d,m}$, while the shape encoder Eq. 8 processes the predicted canonical depth image $C^{d,m}$. Since the pose may occlude parts of the shape, the shape encoder is designed to avoid obstructions and capture the complete shape for accurate reconstruction. The decoder Eq. 10 then reconstructs the voxelized shape ($Y_{shape} \in \mathbb{R}^{256\times256\times256}$).

$$z_{shape} = ShapeEncoder(Y^d, Y^m) \tag{8}$$

$$z_{pose} = PoseEncoder(X^d, X^m) \tag{9}$$

$$Y_{shape} = recon(X^d, X^m, Y^d, Y^m) \tag{10}$$

To smooth the volume surface, we incorporate WGAN-GP in the architecture [9]. We customise the discriminator output to produce a vector rather than a scalar, which stabilises training and improves the quality of the generated surfaces.

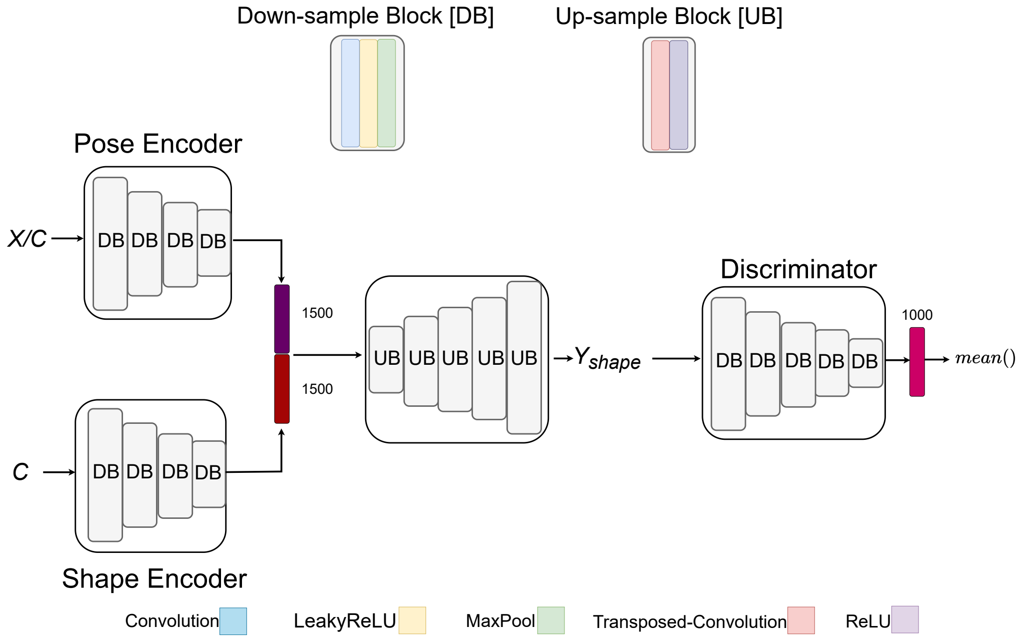

Figure 5: Stage two, in this stage we employ both the original input $X^{d,m}$ and the estimated depth image $C^{d,m}$ to reconstruct 3D shape (Y_{shape}).

3.5. Loss Function

The model is divided into two stages during training. **Stage One** focuses on reconstructing the default pose depth image, while **Stage Two** is dedicated to reconstructing the original pose in the 3D space.

Stage One: This stage employs two loss functions: depth loss and mask loss.

Depth Loss. We use Mean Squared Error (MSE) for the depth loss, modified to concentrate on the foreground region.

$$L_{Depth} = \frac{1}{N} \sum_{i=1}^{N} \hat{y}_m y_m (\hat{y}_d - y_d)^2$$

Here, $\hat{y}_m$ and y_m denote the predicted mask and the ground truth mask, respectively. Likewise, $\hat{y}_d$ and y_d represent the predicted depth and the ground truth depth, respectively. By leveraging the intersection of the masks, we can exclude the background from the depth image, thereby reducing false positive predictions.

Mask Loss. For depth image reconstruction, we desire the model to concentrate on the target shape. Consequently, we aim for the model to learn the canonical form mask.

$$L_{Mask} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_m - y_m)^2 \tag{11}$$

Combined stage one (S1) loss. Since the model has two objectives, we introduce coefficients γ and β to balance the training.

$$L_{S1} = \gamma L_{Depth} + \beta L_{Mask}$$

Stage two: employs two loss function: reconstruction loss and GAN loss.

Reconstruction Loss. We use Binary Cross-Entropy (BCE) as the loss function. However, empty voxels dominate the reconstruction, leading to false negatives. To address this, we add weights to balance the learning process. The modified BCE is shown below.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[-\bar{y}_i \log(y_i) - \alpha(1-\bar{y}_i)\log(1-y_i)].$$

The α is the cost weight on the terms, y and $\bar{y}$ are the estimated shape and ground truth shape respectively.

GAN loss. L_G (Eq. 12) is the loss for the fake estimation, while L_D (Eq. 13) is the discriminator loss used by WGAN-GP [21]. y represents the reconstructed shape and $\bar{y}$ is the ground truth shape. In order to tackle vanishing gradients a weight is introduced (λ) that pushes the gradient norm of the discriminator to be close to 1.

$$L_G = -E[D(y|x)]. \tag{12}$$

$$L_D = E[D(y|x)] - E[D(\bar{y}|x)] + \lambda E[(\|\nabla_{\hat{y}} D(\hat{y}|x)\|_2 - 1)^2].$$

Combined stage two (S2) loss. As the generator has two objectives, a weight is applied to balance both losses during optimisation as follows:

$$L_{S2} = \gamma L_{BCE} + (1-\gamma)L_G. \tag{13}$$

L_{S2} is minimised when training the generator, and L_D is minimised when training the discriminator.

4. Experiments

4.1. Training Details

We split the non-rigid reconstruction task into two stages: **Stage One**, which focuses on reconstructing the canonical pose of a non-rigid object in a depth image, and **Stage Two**, which reconstructs the volume shape using both the original input depth image and the reconstructed depth image (in the canonical pose).

Stage One. The model was trained for 800 epochs. In the initial phase, specifically for the first 100 epochs, we prioritized mask learning. As mentioned in 4.1, depth images are sensitive to mask intersections; therefore, we set $\alpha = 10$ and $\beta = 1000$. In the next 200 epochs, we shifted focus toward the depth objective, setting both α and β to 1000. For the remaining epochs, we allowed the model to concentrate primarily on the depth objective by setting $\alpha = 1000$ and $\beta = 100$. The learning rate was set to 0.001, and we used the Adam optimiser [22].

Stage Two. The model was trained for 500 epochs. During this stage, we froze the model parameters from Stage One while training the Stage Two model. We set $\gamma = 0.8$ and the WGAN-GP gradient penalty to $\lambda = 10$. The learning rate was set to 0.001, and we again used the Adam optimiser [22].

4.2. Datasets

We conducted our experiments on three datasets, all of which contain non-rigid shapes. Specifically, the dataset from [11] features real human data. This dataset was constructed using the Civilian American and European Surface Anthropometry Resource (CAESAR) [23], in which point clouds were fit to templates. In total, it comprises 40 subjects, equally split with 20 males and 20 females. Each subject is represented in 10 different poses.

The second dataset, also from [11], is a synthetic human dataset. It was created in a parameterised manner using 3D modelling software to control the shape and generate poses. This dataset contains 300 shapes, distributed among 15 subjects: 5 males, 5 females, and 5 children. Each subject has 20 poses.

While the aforementioned datasets focus on humans, real-life scenarios present a variety of non-human, non-rigid subjects. As such, we also chose the TOSCA dataset [10], which includes both humans and animals. In total, the dataset has 80 objects. Due to the varied nature of animals, the numbers of poses differ across objects: two males with 7 and 20 poses respectively;

one female with 12 poses; one cat with 11 poses; one dog with 9 poses; one wolf with 3 poses; a horse with 8 poses; a centaur with 6 poses; and one gorilla with 4 poses.

For all datasets, the generation process is as follows: Each shape within the datasets is centred, after which we render an image of size 500 × 500. However, for the TOSCA dataset [10], the sizes of the shapes vary across classes, such as horses and cats. To address this, we scale the shapes to a fixed size (bounding box). We used blender for the dataset generation as we can bind python code to automate the process.

For voxelised ground truth shapes, the datasets offer mesh files which we used for generation.

5. Evaluation

Stage one. For canonical forms, the evaluation measure is typically based on retrieval results [24]. As in previous works [24, 25], we used Clock Matching and Bag-of-Features (CM-BOF) [26] for performing retrieval. The framework starts by computing a descriptor for a given 3D shape. Initially, we centralise the mesh, normalise its scale, and employ a combination of principal component analysis (PCA) and rectilinearity for orientation normalisation. Following this, 66 depth images of the mesh are rendered from 66 viewpoints that are distributed evenly in all directions around a geodesic sphere. Subsequently, SIFT features are extracted from these depth images. Using the bag-of-words method, we generate a histogram descriptor of length 1000 for each image of the shape. The degree of similarity between two shapes is determined by aggregating the similarities of their corresponding views. The retrieval task involves ranking the shapes. For each shape in the dataset, we rank the remaining shapes in relation to it. Once ranked, we employ evaluation metrics to assess the retrieval outcomes. From the literature, we adopt four evaluation metrics: Nearest Neighbour (NN) where the 1-NN algorithm identifies the single nearest neighbour of a query point based on a distance metric (such as Euclidean distance) and assigns the category of this nearest neighbour to the query point. First Tier (FT) refers to a metric that measures the precision at the first rank or the top-n results of the retrieval, assessing how many of the most relevant (or similar) items are correctly identified and ranked by the algorithm at the very top of its output list. Second Tier (ST): while first tier focuses on the precision of the top-ranked results, second tier typically extends this evaluation to a broader set of top results. Discounted Cumulative Gain (DCG) is a measure used to evaluate the effectiveness of ranking algorithms.

Stage two. Treating reconstructed shape and ground truth shape as occupancy volumes, we evaluate our result using Intersection over Union (IoU). The second evaluation metric is mean value Cross-Entropy, Finally, we calculate the Chamfer Distance (CD) by uniformly sampling 30000 points from the surface of both the predicted and ground truth shapes. The resulting CD value is then multiplied by 10^3.

Comparison to prior work.

Stage One. To the best of our knowledge, there exists no learning-based canonical form model specifically tailored for non-rigid shapes. Consequently, all referenced works herein are non learning-based models.

The majority of the methods mentioned in the literature leverage the Multidimensional Scaling (MDS) technique [5]. Hence, (1) MDS-based results are also included in our comparisons. (2) Fast-MDS [27] projects geodesic distances into a Euclidean space. (3) Non Metric MDS, emphasises preserving the ordering of distances rather than their exact values. (4) Least Squares MDS [5], employs the SMACOF (Scaling by Majorising a Convex Function) algorithm. (5) Constrained MDS [28] capitalises on the exact correspondence between an original shape and its Landmark MDS embedding. (6) Detail-preserving Mesh Unfolding method [29] is based on finite elements and omits the use of geodesics. (7) Global Point Signatures (GPS) technique computes the embedding of a mesh. (8) the skeleton based method [30] suggests that a skeleton is derived from a mesh to produce a canonical form. For more details about the previous works, please see the supplementary material.

Stage Two. To evaluate our work, we compare it with methods that perform human reconstruction from a single depth image. Few approaches address this specific task: (1) ShapeFormer, presented by [6], estimates shape from partial point clouds. We extracted point clouds from the depth image to run this model. (2) IF-Net, introduced by [8], provides two different resolution settings for shape completion experiments. Consequently, we used 300-point clouds (IF-Net:300) and 3000-point clouds (IF-Net:3000) for comparison. The method by [7] did not make the code available.

5.1. Results

Stage One. Our model is trained on two datasets and tested on three as stated earlier in Section 4.2. For the synthetic human dataset [11], the results are shown in Table 1. The model is trained using a cross validation method

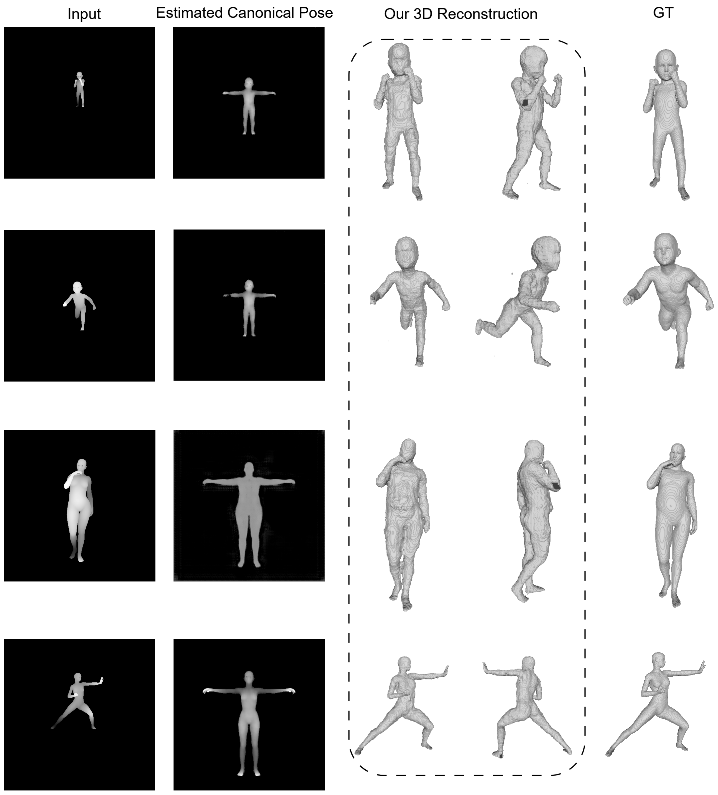

Figure 6: Qualitative results on the synthetic human dataset. The second column shows the estimated canonical pose (Stage 1 result), while the third and fourth columns display the reconstructed shape from different views (Stage 2 result).

where we perform cross validation across the subjects and poses since poses are similar across the whole subjects. Specifically, the subjects and poses are split into groups. Every time, shapes belonging to a chosen group of subjects and a chosen group of poses are used as the test set, while we only use shapes not containing any of these subjects or any of these poses as the training set. This process ensures strict separation of training and test sets during cross validation. The same protocol is applied to other experiments as well.

For the real human dataset [11], the results are shown in Table 2. We trained the model on the synthetic dataset and then tested on the real human dataset. Lastly, for the TOSCA dataset [10] the results are shown in Table 3. In the quantitative results, our model outperforms the state-of-the-art models, except for real human results, our result was the second on the

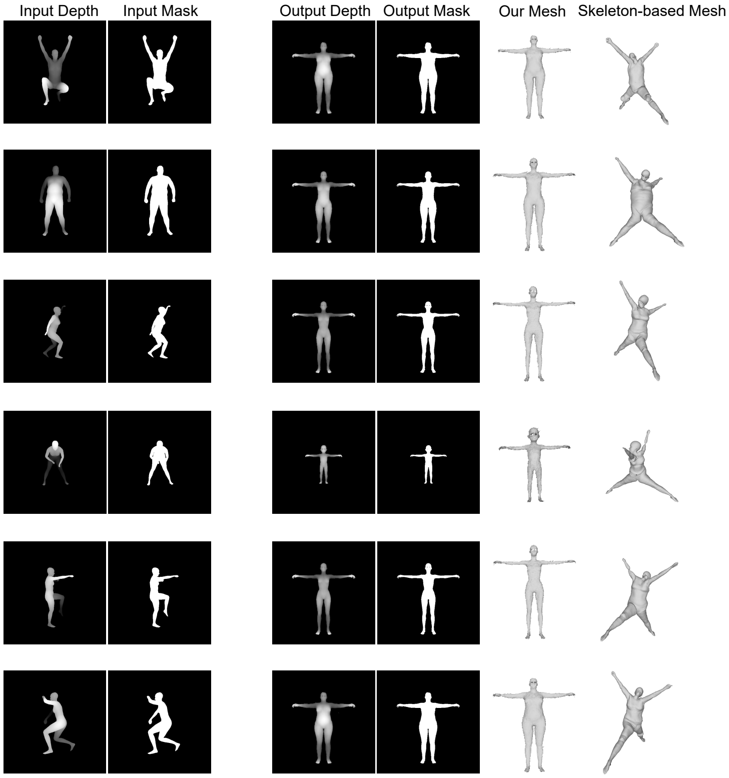

Figure 7: Unseen Canonical form results on real dataset. The model is first trained on synthetic human dataset and then tested on real human dataset. Our meshes are extracted from the output depth images

NN metric, probably due to the domain gap. All the methods performed quite poorly on this dataset, indicating the difficulties for this task. For the qualitative results, for synthetic human dataset [11], the results are shown in the supplementary material, and for the real human dataset [11], the results are shown in Figure 7. For the TOSCA dataset [10] the results are shown in Figure 9.

Stage Two. After training Stage One, we freeze its parameters and train Stage Two. We trained our model on two datasets. For the synthetic human dataset, the results are shown in Table 4. For the TOSCA dataset, the results are shown in Table 5. Our model outperforms the state-of-the-art on both datasets. Results for the qualitative results are shown in Figure 6 and Figure 8.

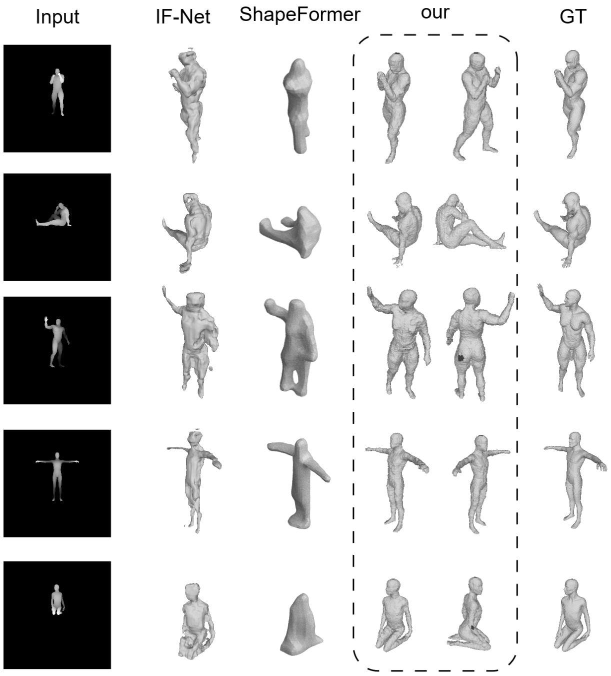

Figure 8: Qualitative results on the synthetic human dataset for Stage two: Pose Recovery in 3D Space.

Generalisability vs. Scalability

Our method is designed with a focus on generalisability rather than scalability. The goal is to achieve robust canonical pose reconstruction from a limited number of training samples, across both human and non-human non-rigid shapes. The model shows strong performance despite being trained on small datasets, demonstrating its potential for broader applicability without relying on large-scale data. While scalability to larger datasets is outside the scope of this work, it remains an important direction for future exploration.

5.2. Ablation Studies

In this section, we conduct two ablation studies using the TOSCA dataset, chosen due to its varied content. Due to space limitations, the qualitative

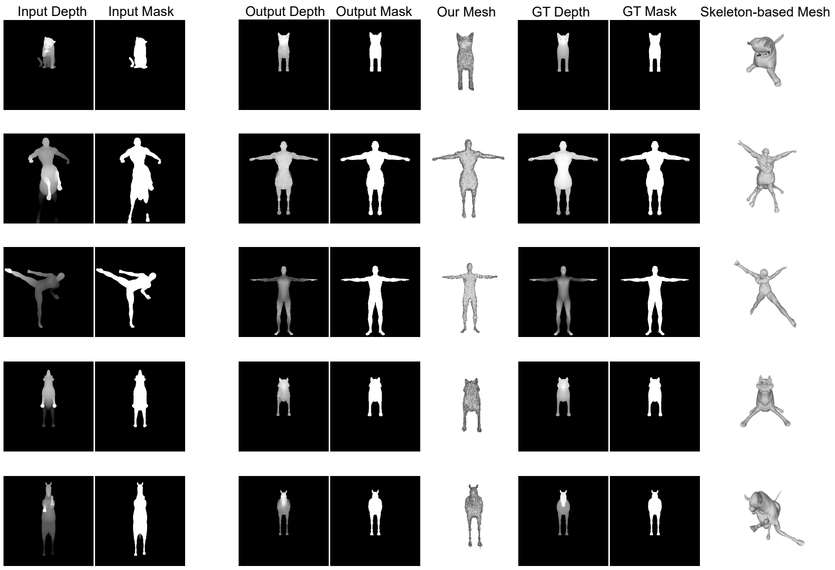

Figure 9: Some canonical form results on the TOSCA dataset. The meshes are extracted from the output depth images

results are presented in the supplementary material.

LFE. Training the model without the LFE component resulted in lower performance compared to the full model. Results are presented in Table 8.

MSFE. Without the MSFE component, the model's performance was worse compared to the complete model (Table 8). For classes like dog or cat (which do not have hand or T-pose features), the model could reconstruct the canonical pose. However, for shapes with outstretched hands and legs, such as centaur or human, the results often missed those body parts.

Shape Encoder. As stated earlier in Section 3.4, the shape encoder helps the model to enhance the reconstruction results. We conducted experiments where the shape encoder was disabled to observe if this led to a degradation in results, see table 6.

Effect of the Discriminator. To evaluate the contribution of the adversarial component, we perform an ablation by removing the discriminator from our model. Without the discriminator, the generator is trained solely with the reconstruction loss, which leads to noisy outputs in the canonical space. These imperfections result in degraded geometric consistency and loss of fine details in the final reconstructions. The discriminator plays a critical role in enforcing global coherence and local surface plausibility, especially

Table 1: Retrieval results for the Synthetic human dataset.

	NN ↑	FT ↑	ST ↑	DCG ↑
Classic MDS	0.10	0.22	0.39	0.54
Fast MDS	0.14	0.20	0.35	0.53
Non-metric MDS	0.09	0.24	**0.41**	0.55
Least Square MDS	0.01	0.13	0.31	0.45
Constrained MDS	0.04	0.14	0.25	0.46
GPS	0.40	0.20	0.32	0.56
Mesh Unfolding	0.04	0.18	0.34	0.49
Skeleton-based	0.01	0.14	0.32	0.46
Our Method	**0.51**	**0.32**	**0.41**	**0.63**

in regions with ambiguous depth input or occlusion. Quantitatively, we observe a notable increase in Chamfer Distance and a reduction in IoU scores when the discriminator is removed, confirming its importance in producing high-fidelity and realistic 3D shapes. The results are shown in Table 7.

6. Conclusion

In conclusion, our research presents a novel learning-based approach that transforms a single depth image into a standard canonical pose (as a depth image) and then recovers the pose in 3D space. Utilising both a depth image and its associated mask, our model successfully estimates the canonical form even for unseen poses. This approach not only aligns diverse input poses into a unified pose but also extends to accurate shape completion in 3D space. Our method demonstrates robustness in generalizing across varied poses and achieves high fidelity in reconstructing detailed 3D shapes.

Funding statements

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number NBU-FFR-2025-2133-02

Table 2: Retrieval results for the real human dataset, trained on the synthetic human dataset and tested on the real human dataset

	NN ↑	FT ↑	ST ↑	DCG ↑
Classic MDS	0.01	0.03	0.07	0.28
Fast MDS	0.00	0.02	0.04	0.27
Non-metric MDS	0.02	0.04	0.08	0.30
Least Squares MDS	0.00	0.00	0.01	0.26
Constrained MDS	0.00	0.01	0.03	0.27
GPS	**0.07**	**0.06**	**0.12**	**0.33**
Mesh Unfolding	0.00	0.01	0.03	0.28
Skeleton-based	0.01	0.01	0.02	0.27
Our Method	0.04	0.023	0.051	0.23

Acknowledgment

This work was carried out as part of the Royal Society International Exchange project Computing Canonical Forms from Collections of 3D Models, IES\R3\183109. Additional funding was provided by the Deanship of Scientific Research at Northern Border University, Arar, KSA on project number NBU-FFR-2025-2133-02.

References

[1] B. Bustos, H. Tabia, J. Vandeborre, R. Veltkamp, Coulomb shapes: Using electrostatic forces for deformation-invariant shape representation, in: Proceedings of the 7th Eurographics workshop on 3D Object Retrieval, 2014, pp. 9–15.

[2] H. Haj Mohamed, S. Belaid, W. Naanaa, L. Ben Romdhane, Local commute-time guided MDS for 3D non-rigid object retrieval, Applied Intelligence 48 (2018) 2873–2883.

[3] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Euclidean-distance-based canonical forms for non-rigid 3D shape retrieval, Pattern Recognition 48 (8) (2015) 2500–2512.

[4] Z. Lian, A. Godil, J. Xiao, Feature-preserved 3d canonical form, International Journal of Computer Vision 102 (2013) 221–238.

Table 3: Retrieval results for the TOSCA dataset. trained on TOSCA and tested on TOSCA.

	NN ↑	FT ↑	ST ↑	DCG ↑
Classic MDS	0.74	0.54	0.80	0.80
Fast MDS	0.73	0.52	0.77	0.77
Non-metric MDS	0.76	0.67	0.87	0.85
Least Square MDS	0.79	0.63	0.86	0.84
Constrained MDS	0.88	0.71	**0.89**	**0.89**
GPS	0.71	0.52	0.72	0.76
Mesh Unfolding	0.88	0.65	0.86	0.85
Skeleton-based	0.78	0.62	0.85	0.84
Our Method	**0.91**	**0.76**	0.80	**0.89**

Table 4: IoU, CD and Cross entropy evaluation metric for the Synthetic Dataset.

	IOU ↑	CE ↓	CD ↓
Shapeformer [6]	0.65	0.059	0.137
IF-Net:300 [8]	0.64	0.059	0.163
IF-Net:3000 [8]	0.69	0.057	0.147
Our Method	**0.81**	**0.039**	**0.094**

[5] A. Elad, R. Kimmel, On bending invariant signatures for surfaces, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1285–1295.

[6] X. Yan, L. Lin, N. J. Mitra, D. Lischinski, D. Cohen-Or, H. Huang, ShapeFormer: Transformer-based shape completion via sparse representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 6239–6249.

[7] B. Zhou, D. Meng, J.-S. Franco, E. Boyer, Human body shape completion with implicit shape and flow learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12901–12911.

[8] J. Chibane, T. Alldieck, G. Pons-Moll, Implicit functions in feature space for 3d shape reconstruction and completion, in: Proceedings of

Table 5: IoU and Cross entropy evaluation metric for the TOSCA Dataset.

	IOU ↑	CE ↓	CD ↓
ShapeFormer [6]	0.57	0.066	0.142
IF-Net:300 [8]	0.51	0.068	0.189
IF-Net:3000 [8]	0.56	0.065	0.163
Our Method	**0.68**	**0.058**	**0.121**

Table 6: Ablation study, IoU and Cross entropy evaluation metric for the TOSCA Dataset.

	IOU ↑	CE ↓	CD ↓
Complete	**0.68**	**0.058**	**0.121**
w/o Shape Encoder	0.59	0.067	0.136

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6970–6981.

[9] B. Yang, S. Rosa, A. Markham, N. Trigoni, H. Wen, Dense 3D object reconstruction from a single depth view, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (12) (2019) 2820–2834.

[10] A. M. Bronstein, M. M. Bronstein, R. Kimmel, Numerical Geometry of Non-rigid Shapes, Springer Science & Business Media, 2008.

[11] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, M. Aono, A. B. Hamza, A. Bronstein, M. Bronstein, et al., Shape retrieval of non-rigid 3D human models, International Journal of Computer Vision 120 (2016) 169–193.

[12] Z. Lian, A. Godil, X. Sun, H. Zhang, Non-rigid 3D shape retrieval using multidimensional scaling and bag-of-features, in: 2010 IEEE International Conference on Image Processing, 2010, pp. 3181–3184.

[13] X.-L. Wang, H. Zha, Contour canonical form: An efficient intrinsic embedding approach to matching non-rigid 3D objects, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, 2012, pp. 1–8.

Table 7: Ablation study, IoU, Cross entropy and Chamfer distance evaluation metrics for the TOSCA Dataset, showing the model with and without Discriminator

	IOU ↑	CE ↓	CD ↓
Complete	**0.68**	**0.058**	**0.121**
w/o Discriminator	0.64	0.061	0.130

Table 8: Ablation study of LFE and MSFE on the TOSCA dataset.

	NN ↑	FT ↑	ST ↑	DCG ↑
Complete	**0.91**	**0.76**	**0.80**	**0.89**
w/o LFE	0.88	0.62	0.76	0.84
w/o MSFE	0.72	0.48	0.61	0.73

[14] H. Zeng, Q. Wang, J. Liu, Multi-feature fusion based on multi-view feature and 3D shape feature for non-rigid 3D model retrieval, IEEE Access 7 (2019) 41584–41595.

[15] M. Jribi, F. Ghorbel, A novel canonical form for the registration of non rigid 3D shapes, in: Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015, Proceedings, Part II 16, Springer, 2015, pp. 230–241.

[16] H. Haj Mohamed, S. Belaid, W. Naanaa, Local feature-based 3D canonical form, in: Representations, Analysis and Recognition of Shape and Motion from Imaging Data: 6th International Workshop, RFMI 2016, Tunisia, October 27-29, 2016, Revised Selected Papers 6, Springer, 2017, pp. 3–14.

[17] X. Wang, A. Boukhayma, et al., 3d human shape and pose from a single depth image with deep dense correspondence enabled model fitting, in: Eurographics Posters, 2022.

[18] Z. Dong, C. Guo, et al., PINA: Learning a personalized implicit neural avatar from a single RGB-D video sequence, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[19] P. Li, W. Zheng, et al., PSHuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025.

[20] Y. Xue, B. L. Bhatnagar, et al., NSF: Neural surface fields for human modeling from monocular depth, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[21] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of Wasserstein GANs, in: NIPS, 2017, pp. 5767–5777.

[22] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).

[23] K. Robinette, S. Blackwell, D. Hoeferlin, CAESAR: Civilian American and European Surface Anthropometry Resource (2002).

[24] D. Pickup, J. Liu, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, S. Nie, L. Jin, G. Shamai, et al., An evaluation of canonical forms for non-rigid 3D shape retrieval, Graphical Models 97 (2018) 17–29.

[25] A. Bronstein, M. Bronstein, U. Castellani, B. Falcidieno, A. Fusiello, A. Godil, L. Guibas, I. Kokkinos, Z. Lian, M. Ovsjanikov, et al., SHREC 2010: Robust large-scale shape retrieval benchmark, Proc. 3DOR 5 (4) (2010) 1–8.

[26] Z. Lian, A. Godil, X. Sun, J. Xiao, CM-BOF: visual similarity-based 3D shape retrieval using clock matching and bag-of-features, Machine Vision and Applications 24 (2013) 1685–1704.

[27] C. Faloutsos, K.-I. Lin, FastMap: A fast algorithm for indexing, datamining and visualization of traditional and multimedia datasets, in: Proceedings of the 1995 ACM SIGMOD International Conference on Management of data, 1995, pp. 163–174.

[28] Y. Sahillioğlu, A shape deformation algorithm for constrained multidimensional scaling, Computers & Graphics 53 (2015) 156–165.

[29] Y. Sahillioğlu, L. Kavan, Detail-preserving mesh unfolding for nonrigid shape retrieval, ACM Transactions on Graphics (TOG) 35 (3) (2016) 1–11.

[30] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Skeleton-based canonical forms for non-rigid 3D shape retrieval, Computational Visual Media 2 (2016) 231–243.