Deep Learning for Canonical Reconstruction of Deformable Objects from Depth Images in Robotic Vision and SLAM-Aware Perception

Fahd Alhamazani
Faculty of Computing and Information Technology, Department of Computer Sciences,
Northern Border University, Rafha, Saudi Arabia

***Abstract*—Reconstructing the canonical pose of non-rigid objects from arbitrary depth observations is an important problem in robotic vision, particularly for systems that must perceive, track, and interact with deformable objects in dynamic environments. In robotics and SLAM-related perception, depth cameras are widely used to support object recognition, spatial understanding, scene mapping, and motion analysis. However, non-rigid deformation remains challenging because the same object may appear in significantly different poses, making reliable object-level representation and tracking difficult. In this study, we present a deep learning approach for reconstructing the default canonical pose of non-rigid objects from single depth images. The proposed model combines short-range and long-range feature extraction with the original depth input to capture both local geometric details and global structural information. By transforming arbitrary posed observations into a consistent canonical representation, the method supports more stable shape understanding, pose normalization, and object-level perception for robotic systems operating in real-world environments. This is particularly relevant to robotic vision tasks involving human motion analysis, deformable-object tracking, manipulation, and semantic mapping. The model is trained on synthetic human datasets and evaluated on synthetic human, real human, and animal datasets. Experimental results demonstrate improved retrieval accuracy compared with existing methods, showing that the proposed approach can generalize across different non-rigid categories and sensing conditions. These findings highlight the potential of canonical pose reconstruction as a useful component for intelligent robotic perception, depth-based scene interpretation, and SLAM-aware systems that require robust understanding of deformable objects.**

***Keywords*—Depth image; computer vision; feature representation; sensor fusion; object-level SLAM; human–robot interaction**

I. Introduction

Reconstructing the default pose of non-rigid objects from arbitrary poses has become an increasingly important task in fields such as computer vision, computer graphics, medical imaging, and robotics. Non-rigid objects, such as human bodies, animals, and deformable objects, undergo complex deformations that make accurate reconstruction and pose normalization a significant challenge. This task is crucial for numerous applications, including animation, motion tracking and shape analysis. Understanding how to reconstruct a standard or default pose from an arbitrary one is not only vital for precise shape representation but also enhances the ability of intelligent systems to interpret and manipulate 3D objects.

In many real-world applications, obtaining consistent 3D representations from varying poses is necessary. For instance, in animation, characters often need to be represented in a canonical pose to ensure uniformity across different scenes and actions. Similarly, in medical imaging, normalizing poses allows for more consistent comparisons of anatomical structures, which is crucial for accurate diagnostics and surgical planning. In robotics, understanding non-rigid body deformations, such as the movements of animals or humans, helps in designing systems that can interact with or mimic these behaviors effectively.

Existing methods for 3D shape reconstruction have focused on either rigid or non-rigid objects, with varying degrees of success. Traditional rigid body reconstruction methods, such as Principal Component Analysis (PCA) or rigid transformation matrices, are not effective for non-rigid objects, as they cannot handle the complex, localized deformations that occur in flexible bodies. On the other hand, non-rigid reconstruction methods, such as Multidimensional Scaling (MDS) [1], [2], Global Point Signatures (GPS) [3], and detail-preserving mesh unfolding techniques [4], have been designed to capture these deformations but often struggle to balance between global structure preservation and fine-grained local details. Many of these methods require tuning or additional constraints to ensure accurate results, particularly when dealing with large datasets or objects with high variability in pose and shape.

Our work introduces a novel approach for reconstructing the default pose of non-rigid objects using depth images as input. Depth images offer rich geometric information, making them a valuable input format for understanding object deformations. By leveraging both short-range and long-range features, our model can effectively capture fine-grained details, such as edges and textures, while also maintaining the overall structure of the object. This dual approach allows the model to handle complex deformations that occur across varying scales, making it suitable for reconstructing both local and global characteristics of non-rigid objects.

We focus on three main components within our model: short-range feature extraction, long-range feature extraction, and the final reconstruction component. Short-range features help capture detailed deformations in small regions of the object, while long-range features ensure that the model accounts for broader spatial relationships across the entire object. The combination of these components, along with the original input image, enables the model to reconstruct the default pose with

high accuracy. This approach is particularly effective for non-rigid bodies, where both local and global deformations must be accounted for.

In summary, this study presents a new method for reconstructing the default pose of non-rigid objects from arbitrary poses using depth images. By combining short-range and long-range feature extraction, our approach provides a comprehensive solution for capturing both fine details and global structure in non-rigid objects. The results of our experiments show that the model is highly effective in various settings, including human and animal datasets, and generalizes well even when trained on synthetic data. Our findings contribute to the growing body of research in shape analysis and 3D reconstruction, with potential applications in animation, medical imaging, and robotics.

Overall, the main contributions of this study are summarized as follows:

- A short-range feature extraction method to capture local details.
- A long-range feature extraction method to capture global structures.
- The results demonstrate that our model successfully recovers the canonical pose for unseen samples.

II. Related Work

The problem of deforming non-rigid objects to a canonical pose from single depth images has been widely addressed in the literature through various methodologies, including learning-based approaches, shape priors, and deformation networks.

One line of work focuses on Multidimensional Scaling (MDS) approaches. For example, Fast-MDS, introduced by Faloutsos and [1], was an early attempt to reduce the computational overhead associated with MDS, allowing for faster shape matching and reconstruction. Despite its efficiency, Fast-MDS struggles with complex non-rigid deformations, limiting its applicability for canonical pose reconstruction from single depth images. Similarly, Non-Metric MDS and Least Squares MDS, proposed by [2], were designed to handle surface bending and stretching, but these methods also face challenges when dealing with high variability in non-rigid shapes and often require multiple observations to achieve robust results.

Recent advancements have expanded upon these MDS-based methods. Constrained MDS, introduced by [5], integrates additional shape constraints to improve non-rigid shape matching and deformation. This method shows improvements over previous MDS techniques but is limited in its ability to handle single-depth-image inputs, as it still relies on some prior shape knowledge.

In addition to MDS-based methods, shape reconstruction techniques have evolved to focus on the preservation of local and global structures. Global Point Signatures (GPS), combined with skeleton-based matching, was proposed by [3] as an effective approach for non-rigid shape matching. GPS, however, falls short in scenarios where fine-grained deformations and high-resolution details are necessary for accurate canonicalisation.

One of the more recent approaches, Detail-preserving Mesh Unfolding also, developed by [4], focuses on preserving intricate surface details while deforming a shape to its canonical pose. This method has been particularly effective in handling high-resolution surfaces, though it requires significant computational resources, which limits its scalability to real-time applications.

Moreover, learning-based approaches like Deform2NeRF and DIF-Net [6], [7] have shown promising results in deforming objects into canonical poses. These methods leverage neural deformation fields and 2D-3D feature fusion to accurately reconstruct non-rigid objects. While these methods achieve high accuracy in reconstructing complex deformations, their heavy reliance on neural networks introduces high computational costs and data requirements, making them less practical for real-time applications or scenarios with limited training data.

Despite the advancements, the field still faces challenges in developing methods that can robustly handle highly complex, non-rigid deformations with minimal data and in real-time scenarios. Methods that balance computational efficiency with high reconstruction fidelity are crucial for advancing the state-of-the-art in non-rigid shape reconstruction and canonical pose estimation.

III. Methodology

The goal is to restore a non-rigid body to its default pose. The shape, represented in a 2.5D image, may appear in any pose and could be subject to significant deformation, causing parts of the shape to become occluded or heavily distorted. Therefore, the model is designed to learn how to reconstruct the non-rigid shape in a canonical pose, The complete model is illustrated in Fig. 1.

The model consists of three components. The first stage, Section III-A, captures fine-grained, localized information (e.g., edges), which helps in accurately modeling small-scale deformations in parts of the object. This provides the model with crucial details on how local areas deform during movement, essential for recovering fine details in the reconstructed default pose.

The second component, Section III-B, captures spatial relationships over larger distances within the object. These features ensure that parts of the object that are far apart, such as limbs in a human body, remain correctly aligned relative to each other. This step helps maintain the global coherence of the object's structure during reconstruction, ensuring the reconstructed pose is anatomically accurate and free from unnatural distortions.

Finally, the third stage, Section III-C processes the outputs of both the first and second stages to reconstruct the shape in its default pose.

A. Short Range Features

Extracting short range features features from depth images, especially in the context of non-rigid shape deformation help model to focus on the fine-grained information in small regions of the image, such as edges, textures, or surface variations [8]. This helps in identifying detailed characteristics of the shape,

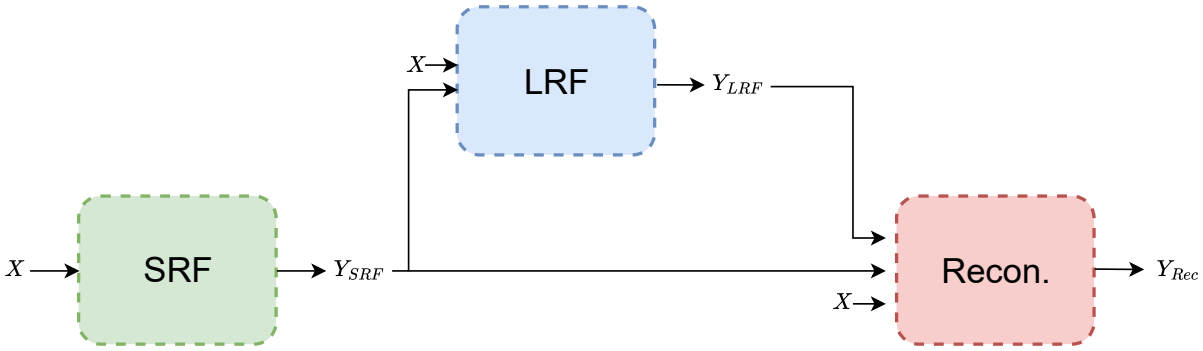

Fig. 1. The complete model first extracts short-range features. Afterward, both the original input and the output of the SRF model (Y_{SRF}) are used as inputs for the long-range features model, which extracts scaled features (Y_{LRF}). Finally, the canonical pose is reconstructed through the reconstruction model (Y_{Rec}).

which are essential for accurately representing and reconstructing the object's geometry in its default pose. Furthermore, Non-rigid objects can deform in complex ways (bending, stretching, etc.). Local features can capture these deformations at small scales, which helps in tracking and understanding how different parts of the shape move or change during deformation (see Fig. 2).

Given a depth image I, where $I \in \mathbb{R}^{500\times500}$, the Short Range Features (SRF) consist of five convolution layers that process the input depth image to extract the short range features:

$$Y_{SRF} = SRF(I)$$

SRE consist of N block, where $N = 9$, each block consist of convolution and Transpose-convolution. At the end, the output Y_{SRF} going to be $Y_{SRF} \in \mathbb{R}^{500\times500}$.

B. Long Range Features

Capture the broader spatial dependencies within the depth image and are crucial for maintaining the overall structure and coherence of the non-rigid object during reconstruction. These features help the model account for how distant parts of the object relate to one another, ensuring that even large-scale deformations are represented accurately. The concatenation of the short-range features with the input image provides a richer representation, enabling the model to leverage both local, fine-grained details and broader spatial information. This holistic understanding ensures that distant parts of the object, such as the extended surfaces, are correctly aligned and positioned relative to one another in the default pose [9] [8].

$$Y_{LRF} = LRF(I, Y_{SRF})$$

Given I and Y_{SRF} input depth image and extracted local features, respectively, the Long Range Feature (LRF) process both the image and the extracted features to produce the long features. The model consist of N block where $N = 5$, each block consist of convolution layer with dilation value to capture long range. The dilation range values in order $dilation = [2, 3, 4, 5, 6]$.

C. Default Pose Reconstruction

The reconstruction component synthesizes all this information to output the default pose in the form of a depth image. The model uses both localized details (from short-range features) and overall structure (from long-range features) and to infer the original, undeformed pose (input depth image) were also utilised. This approach allows the Default Pose Reconstruction (DPR) component to handle complex deformations by integrating both fine-grained and large-scale spatial information, ensuring accurate and detailed default pose reconstruction.

$$Y_{DPR} = DPR(I, Y_{SRF}, Y_{LRF})$$

Given I, Y_{SRF}, and Y_{LRF}, the DPR component processes this information to reconstruct a shape in the default pose, Y_{DPR}, where $Y_{DPR} \in \mathbb{R}^{500\times500}$. The component consists of an Encoder-Decoder architecture, where the encoder is made up of $N = 4$ blocks, and the decoder consists of $K = 4$ blocks. Each encoder block contains convolution, LeakyReLU activation, and max-pooling layers to reduce dimensionality, while each decoder block employs transposed convolution and ReLU activation.

D. Loss Function

The model utilise two loss functions (see Fig. 3):

Depth Loss. We employ the Mean Squared Error for the depth loss:

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

Here, $\hat{y}_d$ and y_d represent the predicted depth and the ground truth depth, respectively.

Mask Loss. As we want the model focus on the shape only, we make the model learn the mask of the shape by adding mask loss:

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

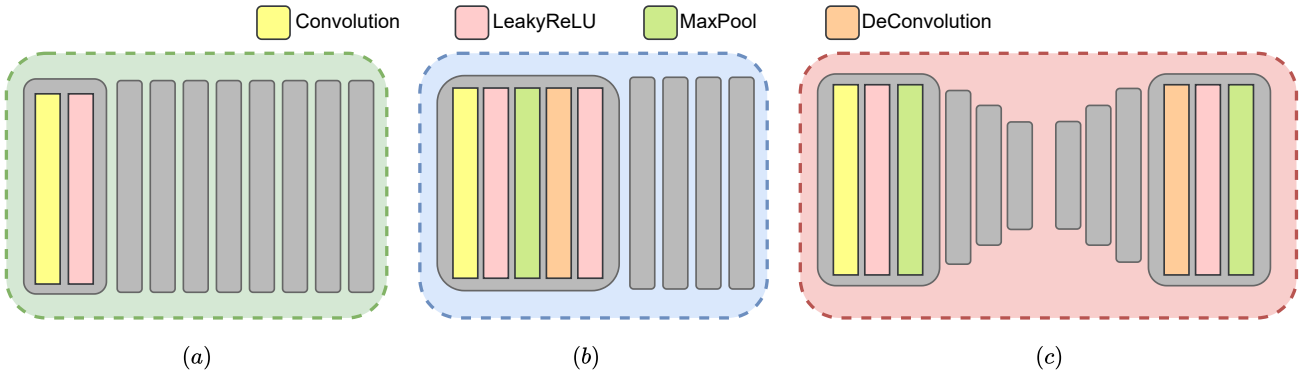

Fig. 2. (a) Shows the SRF layers, where the sub-model consists of nine blocks. Each block contains a convolution layer followed by a LeakyReLU activation. (b) Shows the LRF layers, consisting of five blocks, where each block includes both convolution and deconvolution operations with different dilation rates. (c) Shows the reconstruction sub-model, which includes an encoder–decoder architecture.

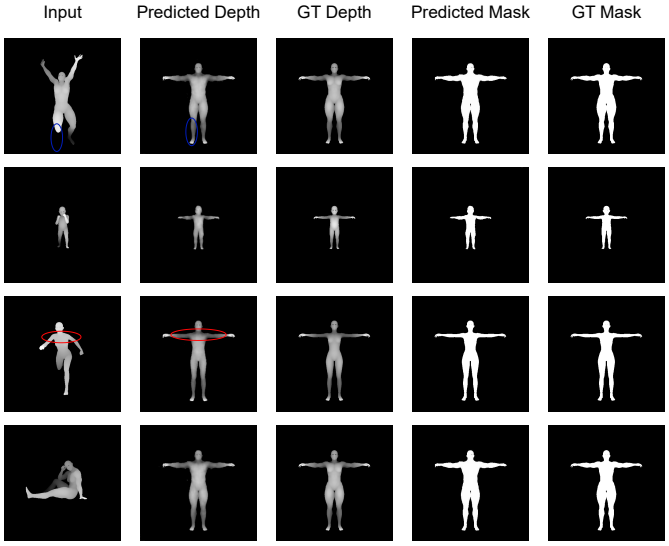

Fig. 3. Our model results on the synthetic dataset. The model predicts both the depth image (second column) and the mask (fourth column). The blue circle shows that the model can recover missing parts. The red circle shows that the model can correctly predict the depth values for the shoulders.

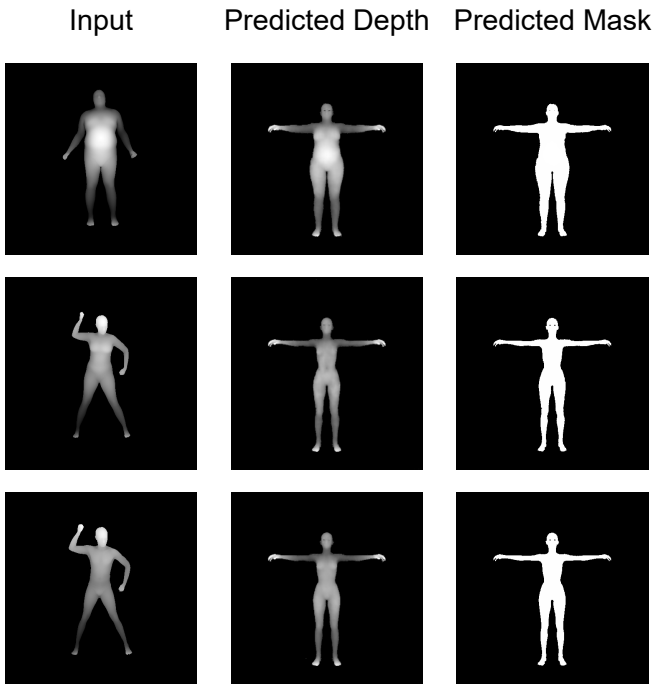

Fig. 4. Our model results on real human datasets. The model was trained on synthetic data and then tested on a real human dataset.

Combined Loss. We introduce coefficients α and β to balance the training:

$$L_{weighted} = \alpha L_{Depth} + \beta L_{Mask}$$

IV. Experiments

A. Training Details

We trained our model for 500 epochs using an end-to-end approach. In the initial phase of training, we set the weighting parameter $\beta = 1$ and $\alpha = 0.01$ for the first 100 epochs. This strategy was employed to allow the model to primarily focus on learning the mask of the shapes during the early stages, thereby reducing the overall complexity of learning the canonical form in the depth format. After the first 100 epochs, both α and β were set to 1, enabling the model to simultaneously learn both the mask and the canonical form with equal emphasis. This progressive learning approach was validated through an ablation study, which demonstrated its effectiveness in simplifying the learning process for depth-based reconstruction.

The model was trained using the Adam optimizer with a learning rate of 0.001, on an NVIDIA RTX 4090 GPU with 24 GB of memory. The total training time was approximately three days with a batch size of 8.

B. Dataset

We conducted our experiments using three datasets: two human datasets and one animal dataset.

The first dataset, provided by [10], consists of 400 samples captured from real human subjects. This dataset represents 40 unique individuals, with each subject contributing multiple samples.

The second dataset, also from [10], comprises synthetic human subjects and includes 300 samples in total. These synthetic samples offer controlled variations that complement the real-world dataset.

To demonstrate the generalizability of our model beyond human subjects, we employed the TOSCA dataset [11], which contains 3D models of animals. This allows us to evaluate the model's ability to reconstruct non-human shapes effectively.

All datasets were processed using Blender and automated with Python scripts. For each sample, we generated depth images of size 500×500, ensuring that the shapes were centered in the frame before capturing the images.

C. Evaluation

For evaluation, we followed the methodology used in previous studies [12], [13], and assessed our model based on retrieval results. Specifically, we employed the Clock Matching and Bag-of-Features (CM-BOF) approach to produce the retrieval results. For a more detailed explanation of CM-BOF, refer to [14].

We selected four evaluation metrics to assess the performance of our model. The first metric is Nearest Neighbour (NN), which finds the nearest neighbor to a point using a specified distance metric. The second metric is First Tier (FT), which measures the precision at the first rank or within the top-n results of the retrieval. The third, Second Tier (ST), extends FT by considering a larger set of result (top-m where $m > n$). Finally, we use Discounted Cumulative Gain (DCG) to evaluate the effectiveness of the ranking algorithm by taking into account the relevance of retrieved items and their positions in the ranking.

D. Comparison to Prior Works

For comparison, we selected several methods from previous studies. These include the Fast-MDS approach proposed by [1], and the Non-Metric MDS method, based on Multidimensional Scaling (MDS), introduced by [2]. Additionally, we included Least Squares MDS from the same study [2].

We also compared our method against the Accelerated MDS approach [15], which improves the efficiency of traditional MDS algorithms. Furthermore, we considered Constrained MDS, a method that integrates specific shape constraints during the MDS process [5].

In addition to MDS-based methods, we evaluated our approach against shape analysis techniques such as Global Point Signatures (GPS), and a Skeleton-based method [3], both of which have demonstrated effectiveness in shape matching tasks. Finally, we included the Detail-preserving Mesh Unfolding method [4], which focuses on preserving fine geometric details during shape reconstruction.

TABLE I. RETRIEVAL RESULTS FOR 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	**0.61**	**0.53**	**0.73**	**0.78**

TABLE II. RETRIEVAL RESULTS FOR REAL HUMAN DATASET, TRAINED ON SYNTHETIC HUMAN DATASET AND TESTED ON 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 Square 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	**0.1**	**0.09**	0.1	**0.41**

E. Result

We evaluated our model on three datasets: synthetic humans, real humans, and animals. For the synthetic human dataset, we split the subjects into a training set of 10 subjects and a test set of 5 subjects (the results shown in Table I). Our model demonstrated a noticeable improvement over previous methods, showing a significant margin of enhancement in the retrieval results.

For the real human dataset (the results shown in Table II and Fig. 4), we followed an unseen setup, where the model was trained on the synthetic dataset and tested on the real human data. Despite the challenge of domain adaptation, our model exhibited substantial improvements, further confirming its generalization capabilities.

Finally, for the animal dataset (the results shown in Table III and Fig. 5), we used the cat and horse samples as the test set, with the remaining animal data used for training. In this case, too, our model outperformed prior approaches, showing marked improvements in retrieval performance compared to existing methods.

V. ABLATION STUDIES

We conducted several experiments to evaluate the effectiveness of different components of our model. The ablation study is designed to demonstrate the impact of each component on the overall performance, the results are shown in Table IV.

A. Disabling Short-Range Features

In this experiment, we disabled the short-range feature extraction component of the model. As a result, the retrieval performance degraded significantly, indicating that short-range features are crucial for capturing fine-grained details necessary for accurate shape reconstruction.

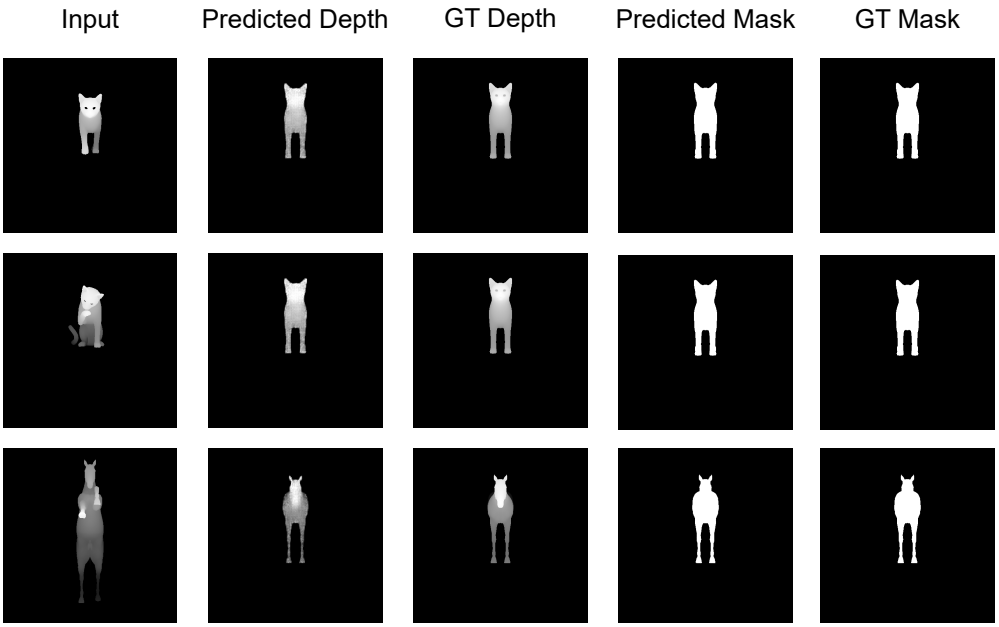

Fig. 5. Our model results on the TOSCA dataset.

TABLE III. RETRIEVAL RESULTS FOR TOSCA DATASET

	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	**0.97**	**0.82**	**0.89**	**0.93**

TABLE IV. RETRIEVAL RESULTS FOR SYNTHETIC DATASET

	NN	FT	ST	DCG
strict train	0.56	0.49	0.66	0.73
w/o SRF	0.58	0.51	0.68	0.76
w/o LRF	0.56	0.50	0.64	0.71
Our	**0.61**	**0.53**	**0.73**	**0.78**

B. Disabling Long-Range Features

When we disabled the long-range feature extraction component, we observed a noticeable degradation in performance compared to the full model. This highlights the importance of long-range features in learning the global structure of the object and suggests that the model effectively utilizes these features to improve reconstruction accuracy.

C. Strict Train-Test Split

To further challenge the model, we performed a strict split, where both pose and subject were excluded from the training set. We implemented this as a cross-validation process, ensuring that the model was tested on all poses and subjects. The model showed resilience under this challenging setup, further validating its generalization ability.

VI. CONCLUSION

In this work, we presented a model for reconstructing the default pose of non-rigid objects from any given pose using depth images as input. Through a combination of short-range and long-range feature extraction, along with the original input, our model demonstrated significant improvements over previous methods, particularly in complex shape deformation scenarios.

Our experiments on synthetic and real human datasets, as well as an animal dataset, showed the model's robustness and generalization capabilities. Specifically, our model achieved notable gains in retrieval accuracy, especially in challenging setups, where the training and test data included unseen poses and subjects.

REFERENCES

[1] C. Faloutsos and K.-I. Lin, "FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets," in *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, 1995, pp. 163–174.

[2] A. Elad and R. Kimmel, "On bending invariant signatures for surfaces," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 10, pp. 1285–1295, 2003.

[3] D. Pickup, X. Sun, P. L. Rosin, and R. R. Martin, "Skeleton-based canonical forms for non-rigid 3D shape retrieval," *Computational visual media*, vol. 2, pp. 231–243, 2016.

[4] Y. Sahillioğlu and L. Kavan, "Detail-preserving mesh unfolding for non-rigid shape retrieval," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 3, pp. 1–11, 2016.

[5] Y. Sahillioğlu, "A shape deformation algorithm for constrained multidimensional scaling," *Computers & Graphics*, vol. 53, pp. 156–165, 2015.

[6] Y. Deng, J. Yang, and X. Tong, "Deformed implicit field: Modeling 3d shapes with learned dense correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 286–10 296.

[7] X. Xie, X. Guo, W. Li, J. Liu, and J. Xu, "Deform2nerf: Non-rigid deformation and 2d–3d feature fusion with cross-attention for dynamic human reconstruction," *Electronics*, vol. 12, no. 21, p. 4382, 2023.

[8] D. Liu, X. Liu, and Y. Wu, "Depth reconstruction from single images using a convolutional neural network and a condition random field model," *Sensors*, vol. 18, no. 5, p. 1318, 2018.

[9] Y. Zuo, Q. Wu, Y. Fang, P. An, L. Huang, and Z. Chen, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 297–306, 2019.

[10] 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*, vol. 120, no. 2, pp. 169–193, 2016.

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

[12] 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*, vol. 97, pp. 17–29, 2018.

[13] 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*, vol. 5, no. 4, pp. 1–8, 2010.

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

[15] G. Shamai, M. Zibulevsky, and R. Kimmel, "Accelerating the computation of canonical forms for 3D nonrigid objects using multidimensional scaling," in *Proceedings of the 2015 Eurographics Workshop on 3D Object Retrieval*, 2015, pp. 71–78.