Sequence Modeling and Generative Model Driven Non-Rigid 3D Reconstruction

Yuxin He, Hui Deng, Mingyi He, Yuchao Dai†
School of Electronics and Information, Northwestern Polytechnical University and
Shaanxi Key Laboratory of Information Acquisition and Processing, Xi'an, Shaanxi, 710129, China

***Abstract*—Non-rigid 3D reconstruction is a critical problem in the computer vision community. Although researchers have made significant progress in recent years, two key challenges remain: (1) ensuring temporal consistency in the reconstructed results, and (2) achieving accurate reconstruction under uncertainty in the input data. To address these challenges, we propose a novel framework that integrates a Temporal Convolutional Network (TCN) for modeling sequential dependencies and ensuring temporal consistency, alongside a diffusion-based module that generates pseudo-3D structures to provide teacher supervision. This diffusion-based supervision enhances spatial accuracy, particularly under weakly supervised conditions. Extensive experiments on the Human3.6M dataset show that our approach achieves superior reconstruction performance compared to existing baselines.**

I. Introduction

Non-Rigid 3D reconstruction is one of the main problems in the computer vision community, which takes a set of 2D observations as input and obtains the 3D structure corresponding to them. Unlike rigid reconstruction, where the object maintains a fixed shape, non-rigid reconstruction must handle continuous deformations, which significantly increase the difficulty of accurate 3D estimation. There have been many early approaches to the topic. Kanade *et al*. [1] was the first to propose factorization to model the reconstruction problem as sparse matrix decomposition. Bregler *et al*. [2] then applied factorization to Non-rigid 3D reconstruction. Building on these two work, the researchers have further proposed approaches [3]–[7].

By introducing deep learning techniques, the researchers obtained even more outstanding results [8]–[14].With neural network, non-rigid reconstruction has the advantages of higher accuracy and faster speed compared to traditional mathematical methods. Yet all these methods face the problem that temporal consistency cannot be guaranteed. The single-frame method [9], [10], [12] represented by Novotny *et al*. [8] do not model temporal information and lack the use of contextual information from input data. In contrast, the alignment-based approach [12], [13] represented by Park *et al*. [14] lacks the exploitation of the orderliness of the temporal information.

Temporal modeling has emerged as an essential component to capture dynamic changes over time, suppress prediction noise, and improve stability and realism in reconstructed motion sequences. On the other hand, other methods [11], [15], [16], that make better use of temporal information ignore the loss of reconstruction accuracy due to uncertainty in the input data in complex scenarios. This problem makes reliable non-rigid reconstruction an open research problem.

In this work, we present a novel framework for non-rigid 3D reconstruction that integrates temporal convolutional modeling with diffusion-based pseudo-3D supervision. To address challenges such as temporal inconsistency and the lack of full 3D ground truth, we propose a temporal aggregation strategy using Temporal Convolutional Networks (TCNs), which efficiently capture both short- and long-term motion dynamics via dilated convolutions. In addition, we introduce a diffusion-based teacher model that generates reliable pseudo-3D structures to supervise the learning process, embedding structural priors into the network under weak supervision. Our main contributions are as follows:

- We propose a TCN-based residual correction module that exploits temporal dependencies to refine and smooth single-frame predictions, enhancing temporal coherence.
- We introduce a diffusion-based teacher model that generates pseudo-3D structures to guide network training with limited supervision, significantly improving spatial accuracy and generalization capability.
- Extensive experiments on the Human3.6M dataset demonstrate that our method achieves superior performance in both reconstruction accuracy and temporal stability, advancing the state of the art in monocular non-rigid 3D reconstruction.

II. Related Work

A. Classical Methods for NRSfM

Non-rigid 3D reconstruction has been extensively studied in computer vision. Bregler et al.[2] introduced a matrix factorization framework using shape bases and SVD. Akhter et al.[3] extended this by introducing trajectory basis modeling with temporal smoothness. Torresani et al.[5] proposed layered shape bases and weak perspective assumptions to improve flexibility. Dai et al.[6] developed a prior-free approach by enforcing low-rank constraints on the Gram matrix. Zhu et al.[4] incorporated subspace clustering to handle complex motion. Kumar et al.[17] refined trajectory modeling with statistical constraints. Shi et al.[7] combined matrix factorization with

† Corresponding author (daiyuchao@nwpu.edu.cn)
This research was supported in part by the National Natural Science Foundation of China (62271410, 12150007).

979-8-3315-7206-8/25/$31.00

Procrustes alignment to enhance accuracy in sparse NRSfM settings.

B. Deep Learning Models for NRSfM

Recent deep learning approaches have greatly advanced non-rigid 3D reconstruction by moving beyond traditional matrix factorization. Novotny et al.[8] proposed C3DPO, which learns canonicalized 3D poses from 2D keypoints via a decoupled architecture. Kong and Lucey[9] introduced Deep-NRSfM with robust mechanisms for missing data. Park et al.[14] developed PRN, which applies Procrustes alignment to unify outputs and reduce camera variation. These works integrate geometric constraints with end-to-end learning, improving robustness to occlusion and ambiguity.

C. Generative Models for NRSfM

Generative models have been increasingly applied to address the ambiguity in monocular non-rigid 3D reconstruction. Denoising Diffusion Probabilistic Models (DDPMs)[18] progressively corrupt data with noise and learn to reverse this process, allowing sampling of diverse 3D hypotheses from learned distributions. Building on this idea of generative diversity, Li and Lee[19] utilized Mixture Density Networks (MDNs) to model multiple pose hypotheses from 2D observations. Sharma et al.[20] proposed a CVAE-based approach to sample 3D poses with structural consistency. Wehrbein et al.[21] introduced Normalizing Flows for probabilistic shape modeling, and Li et al.[22] applied GANs to improve 3D realism. MHFormer[23] further integrated multi-hypothesis and self-attention mechanisms to handle occlusion and depth ambiguity.

III. Prerequisite

A. Non-rigid 3D Reconstruction

Firstly, the basic definition of Non-Rigid 3D Reconstruction is introduced, and classical modeling methods are reviewed to facilitate the subsequent explanation of our approach. The problem can be formulated as:

$$\mathbf{W} = \Pi\mathbf{R}\mathbf{S}, \tag{1}$$

where $\Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ is the orthographic projection matrix. The 2D observation matrix $\mathbf{W} \in \mathbb{R}^{2F\times P}$ contains keypoints for P points across \$F\$ frames. Here, $\mathbf{R}$ represents the camera motion, and $\mathbf{S}$ denotes the time-varying 3D shape matrix. Assuming that the shape sequence can be represented using a set of basis shapes, the shape matrix can be factorized as $\mathbf{S} = \mathbf{B}\mathbf{C}$, where $\mathbf{B}$ is the shape basis matrix and $\mathbf{C}$ is the coefficient matrix. Analyzing the structure reveals that $\mathbf{S}$ has a low-rank property.

Dai *et al*. [6] proposed relaxing the rank minimization to a nuclear norm minimization, by reshaping $\mathbf{S}$ to $\mathbf{S}^\sharp \in \mathbb{R}^{F\times 3P}$, yielding a convex surrogate objective:

$$\min_{\mathbf{S}} \|\mathbf{S}^\sharp\|_*, s.t. \mathbf{W} = \mathbf{\Pi R S}, \mathbf{S} \in \mathbb{R}^{3F\times P}, \mathbf{S}^\sharp \in \mathbb{R}^{F\times 3P}, \tag{2}$$

which improves optimization efficiency and convergence.

In contrast, neural network-based approaches formulate the reconstruction as a learning problem. The key components can be expressed as:

$$\mathbf{S'}_i = f_S \odot f(\mathbf{W}_i, \Theta_\mathbf{S}), \mathbf{S'}_i \in \mathbb{R}^{3\times P}, \tag{3}$$

$$\mathbf{R}_i = f_R \odot f(\mathbf{W}_i, \Theta_\mathbf{R}), \mathbf{R}_i \in \mathbb{R}^{3\times 3}, \tag{4}$$

$$\mathbf{S} = f_{aggregate}(\mathbf{S'}), \mathbf{S} \in \mathbb{R}^{3F\times P} \tag{5}$$

where Θ denotes the network parameters.

However, these approaches often rely on dense and accurate keypoint trajectories across long sequences, making them susceptible to noise, occlusions, and missing data. Such limitations hinder their applicability in unconstrained real-world scenarios.

B. Temporal Convolutional Networks

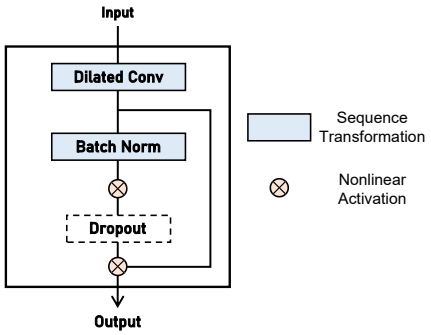

Fig. 1. Structure of TCN Block

Temporal Convolutional Networks (TCN) have emerged as a powerful architecture for sequence modeling, particularly in tasks that require capturing long-range temporal dependencies. TCNs employ causal and dilated convolutions, ensuring that predictions at each time step depend only on current and past inputs. The causal structure respects temporal order, critical for modeling time-series data like human pose sequences. Dilated convolutions allow TCNs to exponentially increase their receptive field without dramatically increasing computational cost, enabling them to model both short-term variations and long-term temporal patterns. Formally, for input sequence $X \in \mathbb{R}^{T\times D}$, the TCN output Y is computed as:

$$Y_t = f(W * X_{t-d:t}), \tag{6}$$

where $*$ denotes convolution with dilation factor d, and f is a non-linear activation such as ReLU. Residual connections are incorporated to stabilize training and allow gradient flow through deep layers, with residual block outputs expressed as:

$$H(X) = \text{ReLU}(X + f(W * X)), \tag{7}$$

which is essential for training deep temporal models efficiently.

IV. Method

Fig. 2 shows the overall framework of our proposed method. It is composed of three interconnected components that collaboratively address the challenges of monocular non-rigid 3D reconstruction. First, the Temporal Convolutional Network (TCN) refines the initial per-frame 3D pose predictions by

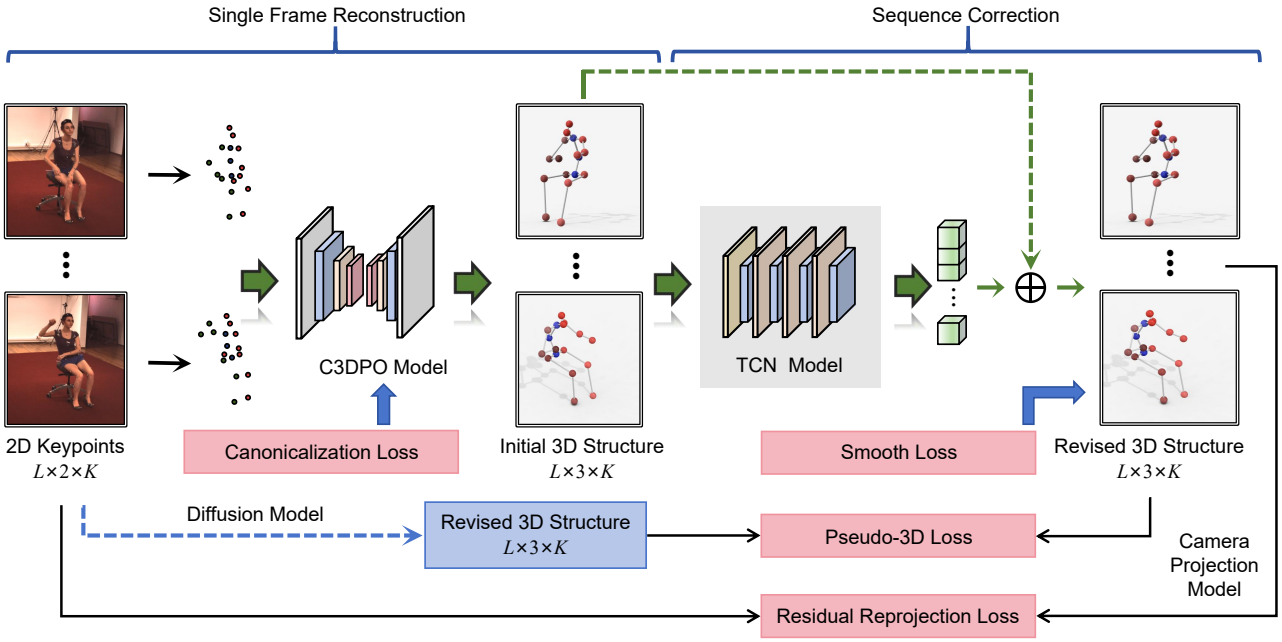

Fig. 2. Overview of our framework and training pipeline. Given input 2D keypoint sequences extracted from monocular images, a Temporal Convolutional Network (TCN) refines per-frame 3D predictions by leveraging temporal context. Simultaneously, a diffusion-based module generates pseudo-3D structures that provide teacher signals. The refined 3D poses are supervised by both residual reprojection loss and pseudo-3D loss, enabling temporally consistent and spatially accurate 3D reconstruction without ground truth 3D supervision.

capturing long-range temporal dependencies, ensuring smoothness and continuity across frames. Second, the diffusion-based generative model serves as a pseudo-3D teacher that provides structural guidance to the network during training. This component enables the network to learn meaningful 3D priors even without direct supervision from ground truth 3D annotations, alleviating the depth ambiguity inherent in monocular settings. Third, the canonicalization and alignment constraint projects predictions into a unified canonical space by learning rotation-invariant representations. This alignment not only improves shape consistency but also stabilizes the learning of pose dynamics. These three modules are tightly coupled: the TCN provides temporally rich features that guide both reconstruction and canonicalization, while the diffusion prior supervises the refinement process and supplies reliable geometric references. Together, they form a cohesive pipeline that delivers robust, temporally consistent, and geometrically accurate 3D reconstructions under weak supervision.

A. Sequence Modeling via TCN

To reconstruct temporally consistent and stable 3D motion sequences for non-rigid objects, we propose a TCN-based architecture that refines single-frame 3D predictions by incorporating rich temporal context. The Temporal Convolutional Network (TCN) backbone consists of four layers with a hidden dimension of 128 channels per layer, enabling effective modeling of long-range temporal dependencies. In our pipeline, the initial per-frame structures $\mathbf{S}_{1:T}$ are processed through a TCN to extract temporal features that effectively capture the dynamics across the entire sequence:

$$\mathbf{F}_{\text{TCN}} = \text{TCN}(\mathbf{S}_{1:T}) \in \mathbb{R}^{T\times D}, \tag{8}$$

where $\mathbf{F}_{\text{TCN}}$ encodes the motion dynamics across T frames with feature dimension D, providing crucial cues to model motion continuity.

For each time step t, these temporal features are combined with single-frame 3D poses through a linear fusion layer, allowing the network to integrate both spatial and temporal information into a unified hidden representation:

$$\mathbf{H}_t = \mathbf{W}_X\mathbf{X}_t + \mathbf{W}_F\mathbf{F}_t + \mathbf{b}, \quad \mathbf{H}_t \in \mathbb{R}^d, \tag{9}$$

where $\mathbf{X}_t$ represents the initial 3D keypoints at frame t, and $\mathbf{H}_t$ encodes the combined spatiotemporal context necessary for refinement.

The hidden features $\mathbf{H}_t$ are then passed through a multi-layer perceptron(MLP) to regress the residual corrections $\Delta\mathbf{S}_t$ for each frame, enabling the network to adjust and refine the initial 3D estimates:

$$\Delta\mathbf{S}_t = \text{MLP}(\mathbf{H}_t), \quad t = 1, \ldots, T, \tag{10}$$

which ensures precise modeling of non-rigid deformations over time. We then obtain the final 3D structure as:

$$\mathbf{S}_t^{\text{final}} = \mathbf{X}_t + \Delta\mathbf{S}_t. \tag{11}$$

Finally, we supervise the reconstruction by minimizing a residual reprojection loss that penalizes the difference between projected refined 3D keypoints and observed 2D ground truth,

thereby enforcing both spatial accuracy and temporal smoothness:

$$\mathcal{L}_{\text{reproj-res}} = \sum_{t=1}^{T} \|\Pi_t \mathbf{R}_t \mathbf{S}_t^{\text{final}} - \mathbf{W}_t\|_\epsilon, \tag{12}$$

where Π_t denotes the camera projection matrix, $\mathbf{R}_t$ represents rotation, $\mathbf{W}_t$ are the observed 2D keypoints and $\|\cdot\|_\epsilon$ denotes Huber loss. This comprehensive design leads to robust non-rigid 3D reconstructions that maintain temporal coherence across challenging motion sequences.

To further promote smoothness and suppress jitter, we apply first- and second-order finite difference constraints over the reconstructed 3D trajectory:

$$\begin{aligned}\mathcal{L}_{\text{smooth}} &= \sum_{t=2}^{T} \|\mathbf{S}_t^{\text{final}} - \mathbf{S}_{t-1}^{\text{final}}\|_\epsilon \\ &+\beta \sum_{t=3}^{T} \|\mathbf{S}_t^{\text{final}} - 2\mathbf{S}_{t-1}^{\text{final}} + \mathbf{S}_{t-2}^{\text{final}}\|_\epsilon\end{aligned} \tag{13}$$

This temporal regularization complements the TCN design and further improves the continuity and realism of non-rigid motion reconstructions.

B. Pseudo-3D Supervision with Diffusion Priors

To overcome the scarcity of ground truth 3D annotations in real-world datasets, we adopt a diffusion probabilistic model (DDPM) [18] to generate plausible pseudo-3D labels, which act as teacher signals during training. Diffusion models learn the data distribution by progressively corrupting data with Gaussian noise in a forward process, and training a neural network to reverse this process to recover the original data.

Formally, the forward process is defined as:

$$q(X_t|X_{t-1}) = \mathcal{N}(X_t; \sqrt{1-\beta_t}X_{t-1}, \beta_t I), \tag{14}$$

where X_t denotes the noisy 3D structure at timestep t, and β_t is a variance schedule controlling the noise level. This process gradually transforms clean 3D data X_0 into pure noise.

The reverse process is learned by a denoising network parameterized by θ, which estimates the posterior of X_{t-1} given X_t:

$$p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \Sigma_\theta(X_t, t)), \tag{15}$$

where μ_θ and Σ_θ are predicted by the network at each timestep. This process enables the generation of pseudo-ground truth 3D structures X_t^{pseudo} from noise.

To guide the model's predictions using these pseudo-3D labels, we define a teacher loss that measures the discrepancy between the refined 3D output $\mathbf{S}_t^{\text{final}}$ and the diffusion-generated structure:

$$\mathcal{L}_{\text{teacher}} = \sum_{t=1}^{T} \|\mathbf{S}_t^{\text{final}} - X_t^{\text{pseudo}}\|, \tag{16}$$

which encourages the model to align its prediction with realistic, temporally coherent 3D poses. We follow the conditional strategy of D3DP [24], where the denoising process is conditioned on the 2D keypoints, further enhancing geometric consistency.

By integrating this teacher-guided loss into the training process, our framework benefits from strong spatial priors learned via diffusion, enabling accurate and temporally stable non-rigid 3D reconstructions even under weak or missing supervision.

C. Canonicalization and Alignment Constraint

We enforce a canonical pose space by aligning per-frame 3D structures with a shared reference via Procrustes transformation:

$$\mathcal{L}_{\text{proc}} = \sum_{i=1}^{F} \min_{\mathbf{R}_i \in SO(3)} \|\mathbf{S}_i - \mathbf{R}_i \bar{\mathbf{S}}\|, \tag{17}$$

which promotes global consistency and mitigates ambiguities across different viewpoints. This alignment constraint encourages the model to disentangle rigid motion from non-rigid deformation.

Inspired by C3DPO [8], we also introduce a canonicalization network Ψ to resolve the inherent ambiguity between rotation and non-rigid deformation. Given a predicted 3D structure $\mathbf{X}$ and a randomly sampled rotation $\mathbf{R} \in SO(3)$, the network is trained with the following consistency loss:

$$\mathcal{L}_{\text{canon}} = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{X}_{:,k} - \Psi(\mathbf{R}\mathbf{X})_{:,k}\|_\epsilon, \tag{18}$$

where K is the number of joints and $\|\cdot\|_\epsilon$ denotes a robust norm such as the Huber loss. This loss enforces that rotating the predicted shape and passing it through Ψ should recover the original structure, thus encouraging the network to consistently factor out the camera viewpoint.

Combining both Procrustes alignment and rotation-equivariant consistency helps establish a stable canonical shape space across frames, improving temporal coherence and structural realism in non-rigid 3D reconstruction.

D. Loss Design

The overall loss is formulated as a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reproj-res}} + \lambda_s \mathcal{L}_{\text{smooth}} + \lambda_t \mathcal{L}_{\text{teacher}} + \lambda_c \mathcal{L}_{\text{canon}}, \tag{19}$$

where λ_s,λ_t,$\lambda_c$are trade-off weights hyperparameters.

V. EXPERIMENT

A. Dataset

We conduct our experiments on the Human3.6M dataset, a widely recognized benchmark for 3D human pose estimation. The dataset provides more than 500 hours of RGB video data paired with 2D and 3D joint annotations, covering 15 action categories captured from four different camera viewpoints. For our framework, we extract sequences of 17 commonly used keypoints to construct the training and evaluation samples.

To form temporal sequences for the model, we apply a sliding window of fixed length L with stride L across the 2D keypoint series, where each segment contains L consecutive

2D poses that serve as model input. This design enables the network to learn dynamic motion patterns and temporal consistency. Following the standard protocol, we split the dataset so that training and validation subjects are mutually exclusive, ensuring there is no overlap of actors between the sets for fair evaluation.

B. Experiment Detail

Our model is trained with a single-stage scheme using sequences of 243 consecutive 2D keypoint frames as input. For optimization, we employ stochastic gradient descent (SGD) with a momentum of 0.9, weight decay of 0.0005, and an initial learning rate of 0.001. The learning rate is scheduled to decay by a factor of 0.1 at epochs 60 and 80, following a total of 120 training epochs.We train the model with a mini-batch size of 32. For the diffusion-based teacher, we follow the architecture and training protocol of D3DP [24]. All experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 2080Ti GPU with 8GB memory, using CUDA version 11.8 and Python 3.10, which ensures efficient training and reproducibility.

C. Quantitative results

We report our quantitative results on the Human3.6M dataset using ground truth 2D keypoints (GT-H36M). We follow the standard evaluation protocol in non-rigid 3D reconstruction and report Mean Per Joint Position Error (MPJPE) and Stress metric, both defined as in [8].As shown in Table I, our method achieves a MPJPE of 64.4mm, significantly outperforming all previous baselines, including PRN [14], PAUL [12], ITES [25], PoseDict [25], and others. Compared to the classical C3DPO [8] and DNRSfM [9] models, which achieve 95.6mm and 109.9mm MPJPE respectively, our approach reduces the error by a large margin.indicating improved stability and robustness in the reconstructed 3D structures. These

TABLE I
COMPARISON ON HUMAN3.6M WITH GROUND TRUTH 2D KEYPOINTS

Methods	MPJPE (mm)	Stress
PRN [14]	86.4	-
PAUL [12]	88.3	-
ITES [25]	77.2	-
PoseDict [25]	85.7	-
C3DPO [8]	95.6	41.5
DNRSfM [9]	109.9	35.9
Seq2Seq [26]	79.8	33.8
MHR [11]	72.1	36.4
Ours	**64.4**	**36.0**

results validate the strength of our temporal-aware architecture and the effectiveness of the diffusion-based teacher prior under weak supervision. The improvement in stress values indicates that our method not only predicts accurate 3D poses but also ensures motion consistency across frames.

D. Ablation Study

In this section, we investigate the contribution of each core module in our proposed framework. All experiments are conducted on the Human3.6M dataset under identical training configurations to ensure a fair comparison.

We construct a series of ablated models by progressively incorporating the smoothness regularization, the Temporal Convolutional Network (TCN), and the pseudo-3D supervision provided by diffusion priors. As shown in Table II, removing any single component results in a noticeable drop in reconstruction accuracy and temporal consistency, underscoring the necessity of each component. Among them, the diffusion-based pseudo-3D supervision contributes the most significant improvement, highlighting its key role in resolving depth ambiguities and guiding the network towards plausible 3D shape predictions. Meanwhile, the TCN module and smooth loss further enhance temporal stability and motion continuity. These findings validate the effectiveness and complementarity of the proposed framework components.

TABLE II
ABLATION STUDIES PERFORMED ON THE HUMAN3.6M DATASET

Method	MPJPE (mm)	Improvement(%)	Stress
C3DPO (baseline)[8]	108.338	-	41.5
Ours (w/ Smooth)	95.318	12.0	36.805
Ours (w/ Smooth+TCN)	83.710	22.7	35.948
Ours (w/ Smooth+TCN+D3DP)	64.446	40.7	35.960

E. Analysis

Despite the competitive results achieved without direct 3D supervision, our method still shows limitations in specific complex actions. Notably, failure cases such as extitWalkingDog and extitPhone reveal structural inconsistencies or deviations in predicted poses. These errors stem from the model's limited capacity to capture abnormal or occluded geometries due to its lack of reliance on precise 3D annotations during training. This underscores the limitations of pseudo-label supervision under challenging conditions.

Future work will focus on improving pseudo-3D supervision quality and enhancing robustness against structural outliers. We also plan to explore weaker or fully unsupervised training regimes to further strengthen the model's generalization ability. Moreover, although our experiments are limited to Human3.6M, the proposed framework is dataset-agnostic, and future work will extend evaluation to more diverse benchmarks to validate generalization.

VI. CONCLUSION

In this work, we propose a novel framework for non-rigid 3D reconstruction, combining temporal residual modeling with diffusion-based pseudo-3D supervision. By leveraging temporal convolutional networks (TCNs) to refine single-frame predictions and incorporating a generative diffusion module as a shape prior, our method achieves coherent and spatially plausible 3D reconstructions without relying on full 3D annotations. This strategy not only improves robustness

under occlusions and depth ambiguities, but also reduces the dependence on dense ground-truth supervision. One limitation of our work lies in its narrow application scope; although the proposed method demonstrates promising results on human pose recovery, the current study focuses only on human body applications, without fully exploring the general potential of NRSfM or diffusion priors. Future research may explore the broader applicability of diffusion priors and temporal modeling in unsupervised or more general 3D reconstruction tasks.

REFERENCES

[1] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," vol. 9, no. 2, pp. 137–154, 1992.

[2] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3d shape from image streams," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, vol. 2, 2000, pp. 690–696.

[3] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade, "Nonrigid structure from motion in trajectory space," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[4] Y. Zhu, D. Huang, F. De La Torre, and S. Lucey, "Complex non-rigid motion 3d reconstruction by union of subspaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1542–1549.

[5] L. Torresani, A. Hertzmann, and C. Bregler, "Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 878–892, 2008.

[6] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," *International Journal of Computer Vision*, vol. 107, pp. 101–122, 2014.

[7] J. Shi, H. Deng, and Y. Dai, "Non-rigid structure-from-motion: Temporally-smooth procrustean alignment and spatially-variant deformation modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 446–21 455.

[8] D. Novotny, N. Ravi, B. Graham, N. Neverova, and A. Vedaldi, "C3dpo: Canonical 3d pose networks for non-rigid structure from motion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7688–7697.

[9] C. Kong and S. Lucey, "Deep non-rigid structure from motion with missing data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4365–4377, 2020.

[10] C. Wang, C.-H. Lin, and S. Lucey, "Deep nrsfm++: Towards 3d reconstruction in the wild," in *2020 International Conference on 3D Vision (3DV)*, 2020, pp. 12–22.

[11] H. Zeng, X. Yu, J. Miao, and Y. Yang, "Mhr-net: Multiple-hypothesis reconstruction of non-rigid shapes from 2d views," 2022, pp. 1–17.

[12] C. Wang and S. Lucey, "Paul: Procrustean autoencoder for unsupervised lifting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 434–443.

[13] H. Zeng, Y. Dai, X. Yu, X. Wang, and Y. Yang, "Pr-rrn: Pairwise-regularized residual-recursive networks for non-rigid structure-from-motion," 2021, pp. 5600–5609.

[14] S. Park, M. Lee, and N. Kwak, "Procrustean regression networks: Learning 3d structure of non-rigid objects from 2d annotations," in *European Conference on Computer Vision*, Springer, 2020, pp. 1–18.

[15] H. Deng, J. Shi, Z. Qin, Y. Zhong, and Y. Dai, "Deep non-rigid structure-from-motion revisited: Canonicalization and sequence modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 2681–2689.

[16] H. Deng, T. Zhang, Y. Dai, J. Shi, Y. Zhong, and H. Li, "Deep non-rigid structure-from-motion: A sequence-to-sequence translation perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 814–10 828, 2024.

[17] S. Kumar, "Non-rigid structure from motion: Prior-free factorization method revisited," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 51–60.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[19] C. Li and G. H. Lee, "Generating multiple hypotheses for 3d human pose estimation with mixture density network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9887–9895.

[20] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain, "Monocular 3d human pose estimation by generation and ordinal ranking," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2325–2334.

[21] T. Wehrbein, M. Rudolph, B. Rosenhahn, and B. Wandt, "Probabilistic monocular 3d human pose estimation with normalizing flows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 199–11 208.

[22] C. Li and G. H. Lee, "Weakly supervised generative network for multiple 3d human pose hypotheses," *arXiv preprint arXiv:2008.05770*, 2020.

[23] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "Mhformer: Multi-hypothesis transformer for 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 147–13 156.

[24] W. Shan, Z. Liu, X. Zhang, *et al.*, "Diffusion-based 3d human pose estimation with multi-hypothesis aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 761–14 771.

[25] C. Xu, S. Chen, M. Li, and Y. Zhang, "Invariant teacher and equivariant student for unsupervised 3d human pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 3013–3021.

[26] H. Deng, T. Zhang, Y. Dai, J. Shi, Y. Zhong, and H. Li, "Deep non-rigid structure-from-motion: A sequence-to-sequence translation perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 814–10 828, 2024.