# PoseFiLM-SDF: Pose-Conditioned Neural Distance Fields for Real-Time Human-Robot Collision Avoidance

**Authors:** Jie Liu, Chencong Jin, Guanxia Dai, Tianmei Sun, Zian Liu, Jie Liu

**Date:** 2026-04-08T08:30:26+00:00

## Abstract

Real-time knowledge of human body spatial occupancy is essential for safe human-robot collaboration (HRC), yet existing methods face a fundamental accuracy-efficiency tradeoff. Voxel-based signed distance fields (SDFs) require prohibitive preprocessing per pose change, while geometric approximations are fast but too coarse for tight safety margins. We present PoseFiLM-SDF, a neural framework that reconstructs human body SDFs from two complementary observations readily available from a 3D depth sensor: 32 skeletal keypoints, which encode both body pose and implicit body size information (via inter-keypoint distances), and 2,048 sparse surface points, which provide precise local geometric constraints on body shape. A dual-encoder architecture processes these modalities through independent pathways, with Feature-wise Linear Modulation (FiLM) conditioning the SDF decoder on the estimated pose. This design enables single-pass inference with no per-instance optimization. Against Trimesh-based ground truth, the framework achieves a 657× speedup (0.86 ms vs. 565.9 ms for 1,000 query points) while maintaining MAE of 0.0088 m and sign accuracy of 90.3% on the full test set. Performance generalizes consistently across four motion capture datasets without retraining. We further validate practical utility through integration with an RRT motion planner, demonstrating collision-free trajectory planning enabled by the 0.0088 m MAE accuracy of the reconstructed distance field.

## Full Text

## Preamble

PoseFiLM-SDF: Pose-Conditioned Neural Distance Fields for Real-Time Human-Robot Collision Avoidance Jie Liu1 , Chencong Jin1 , Guanxia Dai1 , Tianmei Sun2 , and Zian Liu1

Department of Electrical Engineering, Hebei Vocational University of Technology and Engineering, Xingtai, China Department of Mechanical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

## Abstract

—Real-time knowledge of human body spatial occupancy is essential for safe human-robot collaboration (HRC), yet existing methods face a fundamental accuracy-efficiency tradeoff.

Voxel-based signed distance fields (SDFs) require prohibitive preprocessing per pose change, while geometric approximations are fast but too coarse for tight safety margins. We present PoseFiLM-SDF, a neural framework that reconstructs human body SDFs from two complementary observations readily available from a 3D depth sensor: 32 skeletal keypoints, which encode both body pose and implicit body size information (via interkeypoint distances), and 2,048 sparse surface points, which provide precise local geometric constraints on body shape. A dualencoder architecture processes these modalities through independent pathways, with Feature-wise Linear Modulation (FiLM) conditioning the SDF decoder on the estimated pose. This design enables single-pass inference with no per-instance optimization.

Against Trimesh-based ground truth, the framework achieves a 657× speedup (0.86 ms vs. 565.9 ms for 1,000 query points) while maintaining MAE of 0.0088 m and sign accuracy of 90.3% on the full test set. Performance generalizes consistently across four motion capture datasets without retraining. We further validate practical utility through integration with an RRT motion planner, demonstrating collision-free trajectory planning enabled by the 0.0088 m MAE accuracy of the reconstructed distance field.

Index Terms—Human-robot collaboration, signed distance fields, neural implicit representations, collision avoidance, motion planning, real-time systems.

I. I NTRODUCTION Safe operation in environments shared with humans demands real-time knowledge of where the human body is in 3D space—not merely what pose the person is in. A motion planner needs to query the signed distance from each robot link to the human body surface thousands of times per planning cycle, and skeletal parameters alone cannot answer such queries without additional geometric computation. International safety standards such as ISO 10218-1 [1] require continuous collision monitoring within each robot control cycle, placing strict latency constraints on any distance field computation.

Existing approaches face a fundamental accuracy-efficiency tradeoff. Voxel-based SDFs [2] require complete recomputation for each pose change—32.2 s for a 323 grid—and suffer from discretization errors that limit accuracy at low resolutions, making them impractical for dynamic HRC scenarios.

Manuscript received Month Day, Year; revised Month Day, Year.

Bounding volume hierarchies [3] support fast queries but produce overly conservative distance estimates due to coarse geometric approximation, unnecessarily restricting workspace efficiency. Neural implicit representations [4], [5] offer continuous, high-quality SDFs, but prior methods either require per-instance latent code optimization at test time [4] or dense point cloud inputs that are impractical to acquire in real-time settings [6], making them unsuitable for real-time applications.

This paper introduces PoseFiLM-SDF, a neural framework that reconstructs human body SDFs from two types of observations naturally available from a 3D depth sensor: skeletal keypoints and sparse surface points. These modalities are complementary—keypoints encode pose and body size, while surface points provide local geometric constraints. Our dualencoder architecture processes them through independent pathways, with Feature-wise Linear Modulation (FiLM) conditioning the SDF decoder on the pose estimate, enabling accurate reconstruction across diverse poses in a single forward pass without per-instance optimization. PoseFiLM-SDF achieves 0.86 ms inference for 1,000 query points (657× speedup) with MAE of 0.0088 m and 90.3% sign accuracy, sufficient for collision-free trajectory generation in dynamic HRC.

The main contributions of this work are as follows. We present a dual-encoder architecture that separates pose and geometric processing, with FiLM conditioning enabling poseaware SDF reconstruction in a single forward pass. We demonstrate that 2,048 sparse surface points combined with 32 skeletal keypoints suffice for accurate SDF generation, challenging the prevailing assumption that dense inputs are necessary.

We provide systematic evaluation across four motion capture datasets demonstrating strong generalization. Finally, we validate practical utility through RRT motion planning integration.

## II. R ELATED W ORK

A. Traditional SDF Generation Voxel-based methods [2] discretize space into a grid and compute distance transforms [7], but suffer from discretization errors at practical resolutions and require complete recomputation per pose change—a fundamental limitation for dynamic HRC. Precomputed voxel SDFs are used in gradient-based planners [8], [9] but assume static environments. Our method

eliminates preprocessing entirely, adapting to pose changes in a single forward pass.

B. Neural Implicit Representations DeepSDF [4] and Occupancy Networks [5] pioneered learning continuous implicit functions for shape representation. At test time, DeepSDF requires per-instance latent code optimization (approximately 100 gradient steps), making total processing time impractical for real-time use. IGR [10] removes ground-truth SDF supervision but shares similar input density requirements. PIFu [11] reconstructs clothed humans from images but operates offline. These methods are not designed for the sparse, dynamic inputs required

in HRC.

C. Neural SDFs for Articulated Bodies NASA [12] represents articulated bodies as collections of per-part neural indicator functions conditioned on bone transformations. LEAP [13] learns articulated occupancy via canonical space mapping with learned linear blend skinning functions. COAP [14] extends LEAP with part-aware local encoders, improving accuracy on highly articulated poses. These methods achieve strong reconstruction quality but share a critical limitation for our setting: they take ground-truth SMPL parameters as direct input, bypassing the sensor observation step. In practice, SMPL parameters must be estimated from sensor data, introducing additional error. Our method operates directly on raw sensor observations—keypoints and surface points—without assuming known parametric pose.

D. Neural SDFs for Robot Perception iSDF [15] performs real-time SDF reconstruction of roomscale environments via online neural field optimization from a stream of posed depth images. It addresses static scene mapping rather than dynamic articulated bodies. ReDSDF [16] proposes regularized deep SDFs for reactive motion generation, demonstrating smooth distance fields for wholebody control, but assumes the body mesh or pose is given and focuses on field smoothness properties rather than sparseinput reconstruction from raw observations. Our work targets the complementary problem: feedforward SDF reconstruction from sparse sensor observations with no per-frame optimization.

E. Collision Detection for Robotics Bounding Volume Hierarchies [3] enable fast queries but produce overly conservative distance estimates due to coarse geometric approximation, unnecessarily restricting workspace efficiency. Our method provides accurate continuous distances in real-time (MAE 0.0088 m), adapting instantly to dynamic human motion without the approximation errors inherent in geometric surrogates.

## III. P OSE F I LM-SDF M ETHOD

A. Problem Formulation Given 32 skeletal keypoints K = {k1 , . . . , k32 } R32$×$3 and 2,048 sparse surface points S = {s1 , . . . , sN } RN $×$3

(with N = 2048), we learn a continuous SDF f : R3 → R.

For query point q R3 : +d(q, $\partial\Omega$) if q outside body f (q) = −d(q, $\partial\Omega$) if q inside body where $\Omega$ denotes the human body volume, $\partial\Omega$ its surface, and d( $\cdot$ , $\cdot$ ) is Euclidean distance. The goal is accurate SDF reconstruction in a single forward pass with no per-instance optimization, enabling sub-millisecond inference.

B. Network Architecture The framework uses a dual-encoder architecture with a FiLM-conditioned decoder (Fig. 1 [Figure 1: see original paper]).

1) Design Rationale: Surface points and skeletal keypoints are complementary: surface points provide precise local geometric constraints on body shape, while keypoints encode both pose configuration and implicit body

size cues (interkeypoint distances carry limb lengths and proportions). A shared encoder would conflate these roles; our dual-encoder keeps them separate. Keypoints appear twice by design: first fused with surface features to produce the shape encoding z, then re-concatenated explicitly into conditioning vector c to preserve the raw pose signal for FiLM modulation —ensuring fine-grained pose variations drive each decoder layer without being absorbed into the shape code.

2) Architecture Details: Surface Point Encoder: We implement a lightweight PointNet with three 1D convolutional layers ($3 \to 32$, $32 \to 64$, $64 \to 64$), each incorporating Batch Normalization and ReLU. Max-pooling aggregates point-wise features into global shape descriptor $\phi$surf R64 .

Keypoint Encoder: Skeletal keypoints K are flattened into a 96-dimensional vector, processed by MLP (96$→128→64→$32) with LayerNorm and ReLU activations, producing $\phi$kpt R32 , a compact encoding of both pose configuration and body size cues.

Feature Fusion: $\phi$surf and $\phi$kpt are concatenated and processed through MLP (96$→128→$64) producing z R64 .

The conditioning vector is c = [z; $\phi$kpt ] R96 : z = MLPf usion ([$\phi$surf ; $\phi$kpt ]) R64 c = [z; $\phi$kpt ] R

FiLM-Conditioned SDF Decoder: The decoder accepts query point q and conditioning vector c to predict signed distance. Our architecture comprises: (1) input projection layer mapping q R3 to R128 with LayerNorm and ReLU, (2) 5 FiLM-modulated hidden layers (each 128$→128), $and(3)outputhead(128→64→$1). Each FiLM layer computes: hl = $\gamma$ l LayerNorm(FCl (hl−1 )) + $\beta$ l

where modulation parameters ($\gamma$ l , $\beta$ l ) R128 for all 5 layers are generated from c via a shared FiLM generator: twolayer MLP (96$→256→1280)$withReLU, producing$1, 280totalparameters(12
The complete model contains 2.8M parameters—a conscious balance between capacity and efficiency.

Dual Encoder

Feature Fusion

FiLM-Conditioned SDF Decoder

Surface Point Encoder

Keypoint Encoder

MLP through independent pathways. FiLM conditioning adapts the SDF decoder to the current pose and body configuration, enabling accurate reconstruction across diverse body shapes and poses.

C. Training Strategy

1) Data Generation: Our training employs the AMASS dataset [18], specifically its CMU subset comprising 2,088 motion sequences. For each motion sequence, we sample multiple frames and generate an SMPL-H mesh [19], [20] with 6,890 vertices. To enhance shape diversity, we perturb the body shape parameters $\beta$: for each frame we create five variants—the original betas plus four random variations sampled from N (0,

0.

2) clipped to [−3, 3]. From each resulting mesh, we extract 32 skeletal keypoints at fixed joint indices (body joints 0–21 and hand landmarks 25, 28, 31, 34, 36 for left; 40, 43, 46, 49, 51 for right) and sample 2,048 surface points via area-weighted sampling. This yields approximately 104,400 training samples in total. The dataset is split into training (90%), validation (5%), and test (5%) sets with no overlapping motion sequences to prevent data leakage.

3) Stratified Query Point Sampling: Uniform sampling yields fewer than 1% of points near the surface, which is insufficient for learning accurate zero-level-set geometry critical for collision checking. We use a three-stratum strategy: 50% near-surface points (surface vertices perturbed along face normals by $\epsilon$ U (−0.03,

0.

3) m), 20% mid-range points (offset $\delta$ U(−0.25,

0.

25) m from surface), and 30% uniform points sampled within an expanded bounding box (original bounds + 0.50 m margin). This achieves approximately 15–18% of query points within 1 cm of the surface, substantially improving learning of the critical zero-level set.

26) Loss Functions:

Ltotal = wsdf Lsdf + wsign Lsign + wch Lchamf er + wz Lreg

SDF Reconstruction Loss (Lsdf ): Mean squared error over P query points: P1 i=1 (f (qi ) − fGT (qi ))2 .

Sign Consistency Loss (Lsign ): Standard MSE loss provides insufficient gradient signal for correct inside/outside classification near the surface, where |fGT | is small. We address this with a scaled binary cross-entropy loss applied to near-surface points N = {qi : |fGT (qi )| < 0.05 m}:

Lsign = BCElogits (10 · f (qi ), 1[fGT (qi ) > 0]) ,

qi N where BCElogits denotes binary cross-entropy with logits (F.binary_{{{cross}}*{{entropy}}*}{{{with}}_{ in PyTorch), which accepts unbounded real-valued inputs and applies the sigmoid internally—resolving the range constraint.

The binary label 1[fGT (qi ) > 0]  {0, 1} indicates whether the query point lies outside the body surface. The scaling factor of 10 amplifies the gradient signal near the zero level set, where sign errors are most consequential for collision checking, without causing saturation in far-field regions.

Chamfer Distance Loss (Lchamf er ): An auxiliary point cloud decoder (64 → 128 → 256 → 768, outputting 256 × 3 points) reconstructs the surface from latent code

z. Symmetric Chamfer distance between decoded and ground truth surface points provides an additional geometric constraint improving overall shape coherence.

Latent Regularization (Lreg ): Lreg =  z 2 prevents overfitting and encourages compact representations. Loss weights: wsdf = 1.0, wsign = 0.1, wch = 0.5, wz =

0.
1.

4) Training Procedure: We train the model using the AdamW optimizer with learning rate $\eta$max = 5$×$10−4 , weight decay $\lambda$ = 10−5 , and betas $ $1 = 0.9, $ $2 =

0.
1. A linear

warmup is applied for the first 5 epochs, followed by cosine annealing to a minimum learning rate $\eta$min = 10−6 . We use a batch size of 128 and apply gradient clipping with max norm 1.0 to stabilize training. To manage memory constraints during training, we store the dataset in compressed NPZ files (50 samples per file) and implement a smart caching strategy with LRU eviction. The model is trained for up to 1000 epochs, with validation performed every 10 epochs on a held-out 5% subset. We employ early stopping with a patience of 100 epochs based on total validation loss, and retain the checkpoint with the lowest validation loss. Training typically converges around epoch 900 and completes in approximately 150 hours on the same hardware.

## IV. E XPERIMENTS

### A. Experimental Setup

1) Datasets: We evaluate PoseFiLM-SDF on four motion capture datasets. The CMU subset of AMASS [18] is used for training, validation, and test splits following the protocol described in Sec. III-C. To assess cross-dataset generalization, we additionally employ GRAB [21], HumanEva [22], and HDM05 [23]. For each of these generalization datasets, we generate test samples following the same procedure as in training, resulting in approximately 2,000 test frames per dataset. No fine-tuning is performed on these datasets.

2) Baselines: Our comparisons include: Trimesh GT (ground truth SDF computation via ray-casting and nearestpoint queries), Voxel 323 & 643 (discrete SDFs via mesh voxelization and Euclidean distance transform), Multi-Capsule (19 capsules with dynamic radii proportional to body height), and AABB (axis-aligned bounding boxes with 0.05m margin).

3) Metrics: We evaluate both SDF Accuracy (MAE, RMSE, Standard Deviation, 95th percentile error, Sign Accuracy measuring correct inside/outside classification) and Computational Performance (query time for various point counts, total pipeline time including preprocessing)—recognizing that both dimensions are critical for practical deployment.

4) Implementation: All models implement in PyTorch, trained/tested on NVIDIA RTX 4070 Ti Super GPU (16GB VRAM). Timing benchmarks employ identical hardware with Intel Core i5-12400F CPU and 32GB RAM using CUDA Events (100 runs, 10 warmup runs) to ensure reliable measurements.

B. SDF Generation Performance Table I compares SDF accuracy across methods on a 100frame subset of the test set.1 PoseFiLM-SDF achieves MAE of 0.0115 m and sign accuracy of 93.5% on this subset, representing a 6.2× MAE improvement over the best geometric approximation (Multi-Capsule: 0.071

m) and substantially higher sign accuracy than all baselines. 1 Voxel methods require 32.2 s (323 ) and 252.6 s (643 ) preprocessing per frame; evaluating on the full test set is therefore computationally prohibitive.

The 100-frame subset is used solely for fair speed comparison with voxel baselines. Full test set results for PoseFiLM-SDF are reported in Table III.

TABLE I: SDF Accuracy Comparison (100-Frame Subset, 1,000 Query Points per Frame)

## Method

MAE↓ RMSE↓

Multi-Capsule Voxel 323 Voxel 643 PoseFiLM-SDF 0.0115

Sign↑

Note: Sign = Sign Accuracy. All methods evaluated on identical 100-frame subset sampled from test set.

Table II presents computational performance. PoseFiLMSDF achieves a 657× speedup over Trimesh GT for 1,000 queries (0.86 ms vs. 565.9 ms) with zero preprocessing overhead. Considering total pipeline time including preprocessing, PoseFiLM-SDF dramatically outperforms voxel methods, which require 32.2 s (323 ) or 252.6 s (643 ) per pose change. point cloud distributions colored by signed distance values.

Ground truth and PoseFiLM-SDF display smooth, continuous distance fields with clear inside/outside boundaries. While voxel methods produce visually similar distributions, their high MAE (0.115 m, Table

I) reflects numerical inaccuracies not apparent in point cloud visualization, and their prohibitive preprocessing cost (32.2 s per frame) makes them impractical for dynamic HRC. Geometric approximations show distinct failure modes: Multi-Capsule overestimates body volume as each capsule is wider than the actual limb it approximates, compressing distance values near the surface; AABB encloses the entire body in a single bounding box, misclassifying large regions outside the body as interior. PoseFiLM-SDF closely matches ground truth in both distance magnitude and spatial smoothness.

Video Demonstration (Part 1): The supplementary video shows dynamic SDF reconstruction over a complete motion sequence from the AMASS dataset, comparing ground truth (Trimesh) and our PoseFiLM-SDF predictions in real-time.

The visualization presents three synchronized views: (1) the original SMPL-H pose with skeletal keypoints, (2) ground truth SDF with query points colored by signed distance, and (3) our PoseFiLM-SDF predictions. The video confirms that our method maintains consistent accuracy across diverse poses including reaching, bending, and articulated arm movements, while achieving sub-millisecond inference.

C. Generalization Performance Table III shows performance across four datasets without retraining. PoseFiLM-SDF achieves consistent results across all datasets: CMU (MAE: 0.0088, sign accuracy: 90.3%), GRAB (MAE: 0.0120, sign accuracy: 87.1%), HumanEva (MAE: 0.0092, sign accuracy: 91.7%), and HDM05 (MAE: 0.0088, sign accuracy: 90.3%). Sign accuracy above 87% across all datasets confirms reliable inside/outside classification for collision avoidance.

TABLE II: Computational Performance Comparison Across Different Query Sizes (100-Frame Subset)

## Method

1K Queries Query Speedup

5K Queries Query Speedup

Trimesh GT Voxel 323 Voxel 643 Multi-Capsule PoseFiLM-SDF

2,989.5 11,711.0 29,440.5 8,235× 8,287× 8,101× 8,227× 0.258 11,587× 11,722× 3,691× 9,444×

6,818× 6,658× 8,202×

20K Queries Query Speedup

50K Queries Query Speedup 6,902× 6,787× 8,784× 9,685×

Note: Query time excludes preprocessing. Voxel methods require significant preprocessing: 32.2s (323 ) and 252.6s (643 ) per frame—a prohibitive cost for dynamic scenarios. Total pipeline time: PoseFiLM-SDF achieves 0.86 ms for 1,000 queries with zero preprocessing overhead, maintaining sub-millisecond inference up to 20,000 queries and scaling sub-linearly beyond that. Timing measured with CUDA Events (100 runs, 10 warmup runs).

1,000 query points colored by signed distance (blue: inside body, red: outside; color range clipped to $\pm 0.10 m for visualization clarity). Top row : Ground Truth, PoseFiLM−SDF (ours), and Voxel $32^{\{3\}}$. Bottom row : Voxel $64\{3\}$$, Multi-Capsule, and AABB.

PoseFiLM-SDF produces smooth, accurate distance distributions closely matching ground truth. Voxel methods appear visually similar but carry high numerical error (MAE 0.115

m) and prohibitive preprocessing cost. Multi-Capsule overestimates body volume due to oversized capsules, while AABB encloses the entire body in a bounding box, both producing large systematic inside/outside classification errors.

TABLE III: Generalization Performance Across Motion Datasets Dataset

RMSE↓

Sign Acc↑

HumanEva HDM05

Average

Note: CMU evaluated on full test set; generalization datasets sampled with 5 beta variations per sequence and 3 frames per variation.

D. Motion Planning Validation We validate PoseFiLM-SDF through integration with an RRT motion planner [24] for a 6-DOF robot manipulator.

The scenario involves a human raising their arm to block the robot' s initial planned path, requiring dynamic replanning. The scenario comprises 10 frames capturing arm-raising motion, spanning approximately 2 seconds of movement. The robot must navigate around the dynamically posed human body while maintaining a safe clearance.

The planner queries PoseFiLM-SDF to check collision status and compute safety margins during tree expansion. Fig. 3 [Figure 3: see original paper] shows a successful trajectory planned around the human' s

TABLE IV: Ablation Study on Full Test Set Reveals Critical Design Choices Model Variant

RMSE↓

Sign Acc↑

Full Model w/o FiLM w/o Keypoint Encoder w/o Stratified Sampling

be inferred from surface points alone. Reverting to uniform query sampling causes MAE to increase to 0.0492 (5.6×) and sign accuracy to drop by 7.8 percentage points, confirming that near-surface concentration is critical for learning accurate zero-level sets.

F. Computational Analysis robot arm (orange) plans collision-free trajectory (green path with waypoints shown as spheres) around human body (beige mesh) raising their arm. Real-time SDF queries enable efficient collision checking during RRT tree expansion with sub-millisecond query performance—demonstrating practical viability for industrial applications.

raised arm. The 0.86 ms query time (encoding + decoding for 1,000 points) ensures SDF evaluation does not bottleneck the planning process.

Video Demonstration (Part 2): The supplementary video demonstrates the complete planning system in a PyBullet simulation environment with a 6-DOF ABB IRB120 robot.

Key features include: (1) Real-time collision monitoring at 30 Hz control frequency, (2) Future collision detection proactively detecting risks 10 steps ahead, (3) Dynamic replanning generating new trajectories in <1 second with Bezier smoothing, (4) Smooth trajectory execution with direct joint control (max 0.012 rad/step), (5) Collision-free motion maintained throughout execution, and (6) Successful goal reaching while navigating around the human obstacle.

These results confirm that PoseFiLM-SDF provides sufficient geometric accuracy for safe motion planning, and that real-time query speed allows integration with standard motion planners without becoming a computational bottleneck.

E. Ablation Study Table IV validates each design choice. Removing FiLM conditioning and replacing it with simple concatenation causes MAE to increase from 0.0088 to 0.0458 (5.2×) and sign accuracy to drop by 17.8 percentage points, confirming that layer-wise multiplicative pose conditioning is essential for accurate pose-dependent geometry learning. Removing the keypoint encoder results in MAE of 0.0281 (3.2× worse) and sign accuracy of 74.9%, demonstrating that the pose and body size information encoded in keypoints is necessary and cannot

Inference timing is measured using CUDA Events (100 runs, 10 warmup). For one frame with 32 keypoints, 2,048 surface points, and 1,000 query points, the breakdown is: PointNet encoding 0.19 ms, keypoint MLP encoding 0.02 ms, feature fusion 0.15 ms, SDF decoding 0.51 ms, totalling 0.86 ms.

Encoding (0.35 ms total) is computed once per frame and can be reused across multiple query batches.

A key advantage is near-constant decoding time independent of query count,

owing to GPU parallelism. As shown in Table II, inference time remains stable from 1,000 to 20,000 queries and scales sub-linearly beyond that, enabling highthroughput distance field evaluation for motion planning.

V. D ISCUSSION

A. Key Findings PoseFiLM-SDF demonstrates that accurate, real-time SDF generation for dynamic human bodies is achievable from sparse sensor observations. The 0.86 ms inference time supports 30 Hz control without bottlenecking the planning pipeline. The ablation results confirm that all three design choices —dual-encoder, FiLM conditioning, and stratified sampling—contribute substantially to performance; removing any one degrades MAE by 3–6×. Strong generalization across four motion capture datasets (average sign accuracy 89.9%) suggests the learned representation captures general human body geometry rather than dataset-specific artifacts.

B. Limitations The current evaluation uses SMPL-H synthetic meshes as ground truth. Real sensor data introduces additional challenges including depth noise, occlusion from clothing, and keypoint estimation error from the pose estimator —all of which affect reconstruction quality. Validation on real depth sensor data (e.g., using the BEHAVE dataset [25]) remains important future work. The method currently assumes a single human in the workspace; extension to multi-person scenarios requires further investigation. Performance also depends on keypoint estimation quality; highly occluded configurations may degrade results.

C. Future Work Key directions include real-sensor validation with datasets such as BEHAVE, incorporation of temporal information for motion prediction, uncertainty quantification for adaptive safety margins, and extension to multiperson scenarios.

VI. C ONCLUSION We presented PoseFiLM-SDF, a neural framework for realtime human body SDF reconstruction from sparse sensor observations. By processing 32 skeletal keypoints and 2,048 surface points through a dual-encoder architecture with FiLM conditioning, the system achieves 0.86 ms inference (657× speedup over exact computation), MAE of 0.0088 m, and 90.3% sign accuracy on the full test set, with consistent generalization across four motion capture datasets. The dualencoder design keeps pose and geometric information in independent processing pathways, with FiLM enabling flexible pose-adaptive decoding. Stratified near-surface query sampling is essential for learning accurate zero-level-set geometry. Motion planning integration confirms that the 0.0088 m MAE accuracy is sufficient for practical human-robot collaboration.

Future work will address real-sensor validation and multiperson scenarios.

[17]

E. Perez et al., “FiLM: Visual reasoning with a general conditioning layer,” in Proc. AAAI, 2018, pp. 3942–3951. [18]

N. Mahmood et al., "AMASS: Archive of motion capture as surface shapes," in Proc. ICCV, 2019, pp. 5442–5451. [19]

M. Loper et al., "SMPL: A skinned multi-person linear model," ACM Trans. Graph., vol. 34, no. 6, pp. 248:1–248:16,

2015. [20]

J. Romero,

D. Tzionas, and

M.

J. Black, "Embodied hands: Modeling and capturing hands and bodies together," ACM Trans. Graph., vol. 36, no. 6, pp. 245:1–245:17,

2017. [21]

O. Taheri et al., "GRAB: A dataset of whole-body human grasping of objects," in Proc. ECCV, 2020, pp. 581–600. [22]

L. Sigal,

A.

O. Balan, and

M.

J. Black, "HumanEva: Synchronized video and motion capture dataset," Int.

J. Comput. Vis., vol. 87, no. 1-2, pp. 4–27,

2010. [23]

M. Müller,

T. Röder,

M. Clausen,

B. Eberhardt,

B. Krüger, and A.

Weber, "Documentation mocap database HDM05," Tech. Rep. CG-20072, Universität Bonn,

2007. [24]

S.

M. LaValle, "Rapidly-exploring random trees: A new tool for path planning," Tech. Rep. 98-11, Iowa State University,

1998. [25]

B.

L. Bhatnagar et al., "BEHAVE: Dataset and method for tracking human object interactions," in Proc. CVPR, 2022, pp. 15935–15946.

## ACKNOWLEDGMENT

This work was supported by funding sources withheld for double-blind review.

R EFERENCES [1] ISO 10218-1:2011, "Robots and robotic devices –Safety requirements for industrial robots –Part 1: Robots," International Organization for Standardization,

2011. [2]

B. Curless and

M. Levoy, "A volumetric method for building complex models from range images," in Proc. SIGGRAPH, 1996, pp. 303–312. [3]

S. Gottschalk,

M.

C. Lin, and

D. Manocha, "OBBTree: A hierarchical structure for rapid interference detection," in Proc. SIGGRAPH, 1996, pp. 171–180. [4]

J.

J. Park et al., "DeepSDF: Learning continuous signed distance functions for shape representation," in Proc. CVPR, 2019, pp. 165–174. [5]

L. Mescheder et al., "Occupancy networks: Learning 3D reconstruction in function space," in Proc. CVPR, 2019, pp. 4460–4470. [6]

J. Chibane,

T. Alldieck, and

G. Pons-Moll, "Implicit functions in feature space for 3D shape reconstruction and completion," in Proc. CVPR, 2020, pp. 6970–6981. [7]

P.

F. Felzenszwalb and

D.

P. Huttenlocher, "Distance transforms of sampled functions," Theory Comput., vol. 8, no. 1, pp. 415–428,

2012. [8]

N. Ratliff et al., "CHOMP: Gradient optimization techniques for efficient motion planning," in Proc. ICRA, 2009, pp. 489–494. [9]

J. Schulman et al., "Finding locally optimal, collision-free trajectories with sequential convex optimization," in Proc. RSS,

2013. [10]

M. Atzmon and

Y. Lipman, "SAL: Sign agnostic learning of shapes from raw data," in Proc. CVPR, 2020, pp. 2565–2574. [11]

S. Saito et al., "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in Proc. ICCV, 2019, pp. 2304–2314. [12]

B. Deng et al., "NASA: Neural articulated shape approximation,"in Proc. ECCV, 2020, pp. 612–628. [13]

M. Mihajlovic et al., "LEAP: Learning articulated occupancy of people,"in Proc. CVPR, 2021, pp. 10461–10471. [14]

M. Mihajlovic et al., "COAP: Compositional articulated occupancy of people," in Proc. CVPR, 2022, pp. 13201–13210. [15]

J. Ortiz et al., "iSDF: Real-time neural signed distance fields for robot perception," in Proc. RSS,

2022. [16]

P. Liu et al., "Regularized deep signed distance fields for reactive motion generation," in Proc. IROS, 2022, pp. 6673–6680.

*Note: Figure translations are in progress. See original paper for figures.*