SCLTRACK: STRUCTURED CORRESPONDENCE LEARNING FOR CATEGORY-LEVEL GARMENT POSE TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Tracking category-level garment pose from partial point clouds requires correspondence to remain reliable despite large non-rigid deformation, self-contact, occlusion, and continuously changing observations. A key difficulty is that dense inter-frame features alone must simultaneously encode regional structure, instantaneous geometry, and persistent identity, although these cues follow different spatial and temporal behaviors. We present SCLTrack, a **S**tructured **C**orrespondence **L**earning framework that explicitly organizes these complementary cues before canonical correspondence estimation. SCLTrack introduces a semantic branch to capture region-level structural context, a geometry branch to preserve local metric relations, and a consistency branch that retrieves temporal evidence from an ordered canonical memory. Their point-wise representations are integrated with the base inter-frame feature through adaptive gated residual fusion. The enhanced representation is further refined in NOCS space and decoded through a canonical volumetric representation and implicit warp field, which also updates the canonical memory for subsequent frames. Extensive experiments demonstrate that SCLTrack significantly enhances the accuracy of garment pose tracking by up to 14.5% over GaPT-DAR, while reducing correspondence, surface, and NOCS errors by 2.1%, 6.5%, and 8.9%, respectively.

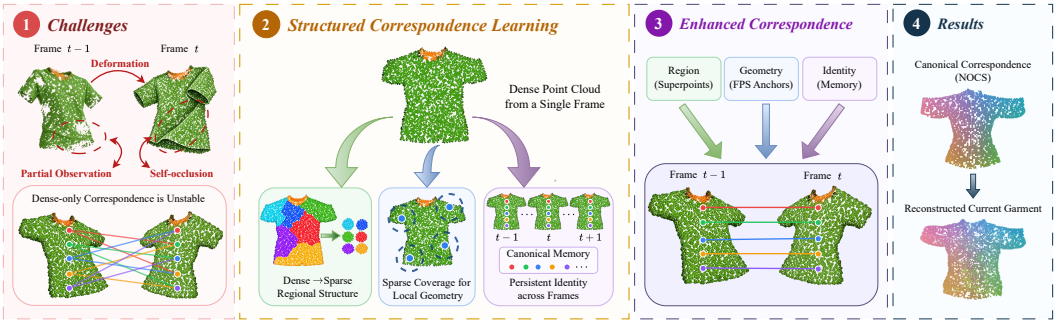

Figure 1: Dense garment correspondence involves a series of challenges for garment pose tracking. We model complementary regional, geometric and identity clues with superpoints, FPS anchors and canonical memory, and adaptively fuse them for robust correspondence learning. This structured representation improves canonical correspondence estimation and garment reconstruction.

1 INTRODUCTION

Garments are highly deformable objects, and embodied systems must perceive their states reliably in order to manipulate, fold, and otherwise interact with cloth. Unlike rigid objects, an off-body garment has effectively infinite degrees of freedom and can undergo large bending, folding, self-contact, and self-occlusion, so that only a partial and rapidly evolving portion of its surface is observable in any given frame (Chi & Song, 2021; Xue et al., 2023; Li et al., 2024). Consequently, category-level methods commonly adopt a shared canonical space to represent unseen instances and to es-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tablish correspondences between partial observations and the complete object surface (Wang et al., 2019; Chi & Song, 2021; Zhu et al., 2025). Tracking, however, makes this correspondence problem considerably harder than single-frame estimation. We identify three interdependent representation challenges as particularly critical. (i) *Structural-semantic ambiguity* arises when garment regions that are far apart in the canonical space become locally similar, or even spatially adjacent, under severe folding and self-contact. (ii) *Geometric fidelity* is difficult to preserve, since thin, multi-layered cloth demands precise local metric cues near folds, boundaries, and contacting layers. (iii) *Temporal identity* can be lost when surface regions are occluded and later reappear, and correspondence errors from earlier frames may accumulate and propagate over time. Robust garment tracking thus requires structural, geometric, and temporal cues to remain complementary, rather than entangled in a single point-wise representation.

Canonical-space methods have made substantial progress toward this goal. GarmentNets (Chi & Song, 2021) casts single-frame garment pose estimation as canonical-space shape completion, inferring complete configurations from partial observations. GarmentTracking (Xue et al., 2023) extends this to videos by fusing adjacent-frame features and using the previous NOCS prediction as a positional cue, while GaPT-DAR (Zhang et al., 2025) makes deformation more explicit through 2D warping followed by depth reconstruction. Yet these methods do not factorize correspondence according to the distinct roles of structural context, local metric geometry, and persistent temporal identity: dense point features must encode both regional structure and fine local geometry, while temporal reasoning depends heavily on earlier predictions. These demands conflict when folds bring unrelated regions into contact, or when occluded regions reappear after earlier prediction errors. This raises a more fundamental question: can garment correspondence be made more robust by explicitly separating ***what forms a coherent structure***, ***where that structure lies in the current geometry***, and ***which canonical region it represents over time***? Neighbouring fields of 3D perception and tracking offer useful ingredients. Superpoint-based methods replace isolated points with compact region-level representations (Hui et al., 2022; Geist et al., 2025; Du et al., 2026), and GAPS (Yu et al., 2025) learns adaptive point-to-superpoint associations that yield more coherent structure than fixed grouping. KPA-Tracker (Liu et al., 2024) shows that sparse ordered keypoints can provide persistent identities in articulated-object tracking. None of these transfers directly to garments, however: adaptive superpoints lack temporal ordering, farthest-point-sampled anchors lack persistent identity, and articulated keypoints assume rigid-body motion rather than the many degrees of freedom of cloth. Garments thus call for a dedicated division of labor: adaptive superpoints for semantic structure, independent geometric anchors for current-frame metric structure, and ordered canonical anchors, updated through a non-rigid warp, for temporal identity.

Motivated by this insight, and as illustrated in Figure 1, we propose **SCLTrack**, a **S**tructured **C**orrespondence **L**earning framework for category-level garment pose tracking that explicitly factorizes correspondence into structural, geometric, and temporal components. Rather than replacing the established canonical-space tracking pipeline, SCLTrack intervenes directly at the point features obtained after inter-frame fusion, and enriches them with three complementary, role-specific representations. The three branches introduce complementary inductive biases for regional structure, local geometry, and temporal correspondence. The *semantic branch* addresses *what forms a coherent structure*. It aggregates dense point features into adaptive superpoints, with point-to-superpoint associations determined jointly by feature correlations and geometric relationships. It then propagates the resulting region-level context back to individual points. Since grouping is guided by both appearance-level features and spatial proximity, points that are adjacent in the current frame but belong to different canonical regions are not indiscriminately merged, which mitigates structural-semantic ambiguity under folding and self-contact. The *geometric branch* addresses *where that structure lies in the current geometry*. It applies farthest point sampling (FPS) to construct spatially distributed anchors that are independent of the semantic superpoints, explicitly aggregates the local relative geometry around each anchor, and interpolates this metric context back to the dense representation. This separate branch emphasizes local metric cues near folds, boundaries, and contacting layers. The *consistency branch* addresses *which canonical region a structure represents over time*. It maintains a compact set of ordered canonical anchors that carry a fixed canonical identity, together with their recent task-space states and features, forming a canonical-indexed memory. Current points retrieve temporal context from this memory. After NOCS prediction, the estimated non-rigid warp maps the same anchors into the current frame and refreshes their states for subsequent tracking. The fixed canonical anchors provide a consistent indexing scheme for temporal feature retrieval, while their task-space positions and features are updated after each frame. Finally, a gated residual fusion

module adaptively combines the three complementary cues with the original inter-frame feature before NOCS prediction, allowing the network to weight structural, geometric, and temporal evidence according to the current deformation state while preserving the original feature pathway.

To sum up, our contributions are threefold: (1) We propose SCLTrack to reformulate category-level clothing pose tracking as structured correspondence learning, identifying structural semantics, local metric geometry and canonical temporal identity as the three complementary factors for robust tracking. Under non-rigid deformation, self-contact, and partial observation, this formulation effectively improves the accuracy of correspondence inference. (2) We achieve this formulation through an adaptive superpoint semantic branch, an FPS-based geometric branch, and a persistently ordered consistency branch. Subsequently, a gated residual fusion mechanism integrates these three clues, while an anchor update mechanism in the consistency branch establishes a temporal feedback loop between continuous frames. (3) Extensive experiments and ablation analyses performed on the VR-Folding benchmark dataset demonstrate that this method achieves a significant performance improvement over state-of-the-art methods in the task of category-level garment pose tracking. Additionally, the complementary roles of these three branches have been confirmed.

2 Related Work

Canonical Garment Tracking. Canonical representations provide a shared reference for category-level pose reasoning. NOCS (Wang et al., 2019) introduced normalized object coordinates, while ANCSH (Li et al., 2020) extended canonicalization to articulated objects. Later methods incorporated structural priors (Chen & Dou, 2021), pose consistency (Lin et al., 2021), and implicit interaction modeling (Jiang et al., 2022). For deformable garments, GarmentNets (Chi & Song, 2021) formulated canonical-space shape completion, CAPTRA (Weng et al., 2021) enabled online category-level tracking, GarmentTracking (Xue et al., 2023) coupled canonical and task-space prediction, and GaPT-DAR (Zhang et al., 2025) modeled deformation through 2D warping and 3D reconstruction. However, regional structure, local geometry, and persistent identity remain largely encoded in a shared correspondence representation.

Structured Spatial Correspondence. PointNet++ (Qi et al., 2017), Point Transformer (Zhao et al., 2021), Stratified Transformer (Lai et al., 2022), Point Transformer V2 (Wu et al., 2022), PointMLP (Ma et al., 2022), and PointNeXt (Qian et al., 2022) improve local and neighborhood-aware point representations. Regional abstractions are explored by SPFormer (Sun et al., 2023) and GAPS (Yu et al., 2025), while Point Transformer V3 (Wu et al., 2024) improves large-scale point organization. DV-Matcher (Chen et al., 2025) further combines geometric and visual priors for non-rigid matching. SCLTrack instead jointly models adaptive regional structure and FPS-based local metric geometry before projecting both cues to dense point-wise features.

Persistent Canonical Correspondence. PointDSC (Bai et al., 2021), GeoTransformer (Qin et al., 2022), RoITr (Yu et al., 2023), and Lepard (Li & Harada, 2022) improve pairwise matching under geometric variation, while FlowStep3D (Kittenplon et al., 2021) and SCOOP (Lang et al., 2023) estimate inter-frame motion. KeyTr (Novotny et al., 2022), BundleSDF (Wen et al., 2023), and KPA-Tracker (Liu et al., 2024) maintain temporal evidence through keypoints or memory. SpatialTrackerV2 (Xiao et al., 2025) integrates 3D point tracking with scene geometry and camera motion, while PointSt3R (Guerrier et al., 2026) models tracking through 3D-grounded correspondence. SCLTrack instead maintains ordered canonical anchors with persistent identity and recurrently updated task-space positions, motion states, and features.

3 Method

3.1 Overview

Problem statement. We consider category-level garment pose tracking from sequential partial point observations. At time t, the model receives the previous and current point clouds $P_{t-1} = \{p_{t-1}^i\}_{i=1}^N$ and $P_t = \{p_t^i\}_{i=1}^N$, the previous per-point canonical coordinates $Y_{t-1} = \{y_{t-1}^i\}_{i=1}^N$, and a raw canonical garment surface $Q_t^{raw} = \{q_t^j\}_{j=1}^M$ represented by mesh NOCS samples. The recurrent canonical memory C_{t-1} is maintained internally across consecutive frames. The goal is to predict the current point NOCS $\widehat{Y}_t$ together with the complete garment surface $\widehat{X}_t$ in task space. We express the tracking problem as $(\widehat{Y}_t, \widehat{X}_t) = \mathcal{F}\big(P_{t-1}, P_t, Y_{t-1}, Q_t^{raw}; C_{t-1}\big),\ \widehat{X}_t = W_t(\widehat{Q}_t; V_t)$, where $\widehat{Q}_t$

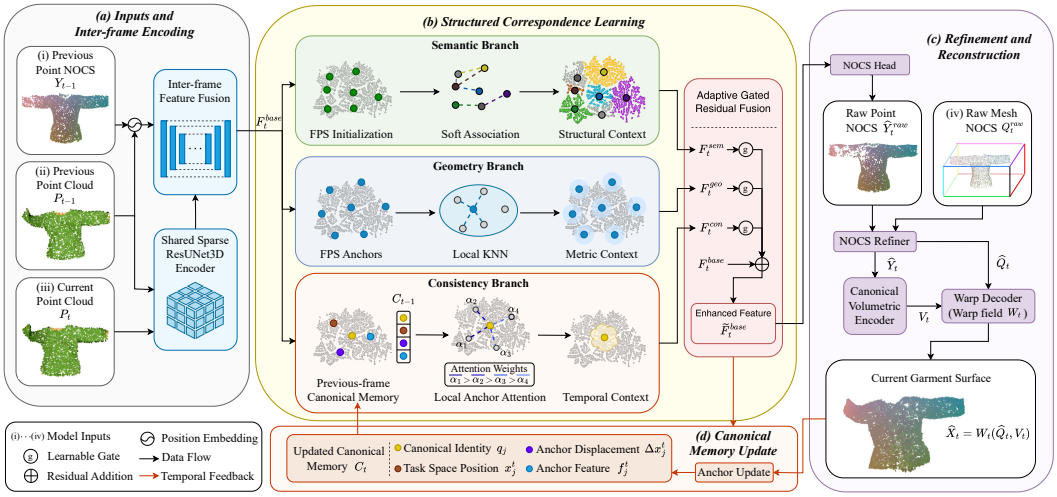

Figure 2: **Overview of SCLTrack.** (a) Adjacent point clouds and previous point NOCS are fused as input to obtain the base representation F_t^{base}. (b) Semantic, geometric, and consistency clues are jointly enhanced by adaptive gated residual fusion to obtain $\widetilde{F}_t^{base}$. (c) Combining NOCS refinement and implicit warping reconstruction to obtain the current garment surface $\widehat{X}_t$. (d) The predicted warp updates the ordered canonical anchors and canonical memory C_t for the next frame.

is the refined canonical surface, V_t is the canonical feature volume, and W_t denotes the learned implicit canonical-to-task warp.

Overall framework. As illustrated in Figure 2, SCLTrack consists of four connected components. (a) Adjacent partial observations are encoded by a shared sparse 3D backbone and fused across frames. (b) The fused point representation is decomposed into structural, local geometric, and canonical-temporal clues through three complementary correspondence branches. Their outputs are adaptively integrated with the original inter-frame feature through gated residual fusion. (c) The enhanced representation predicts current point NOCS and jointly refines the point correspondence and canonical garment surface before reconstructing the complete garment pose through canonical volumetric encoding and implicit warping. (d) The predicted warp refreshes a persistent set of ordered canonical anchors, forming the memory used by the next frame.

3.2 Inputs and Inter-frame Encoding

We independently voxelize the consecutive partial point clouds P_{t-1} and P_t and process them with a shared SparseResUNet3D encoder $\mathcal{E}_{sp}$, obtaining $F_\tau^{sp} = \mathcal{E}_{sp}(P_\tau)$ for $\tau \in \{t-1, t\}$. The inter-frame Transformer $\mathcal{T}$ uses asymmetric positional inputs. The previous frame combines task-space coordinates and point NOCS as $[P_{t-1}, Y_{t-1}]$, while the current frame uses P_t. These six- and three-dimensional inputs associate the current observation with the preceding canonical correspondence, yielding $F_t^{base} = \mathcal{T}\big(F_{t-1}^{sp}, [P_{t-1}, Y_{t-1}], F_t^{sp}, P_t\big)$. The fused representation remains aligned with P_t and provides the common input (P_t, F_t^{base}) to the structured branches in the following section.

3.3 Structured Correspondence Learning

We organize correspondence reasoning around three complementary requirements. *Regional coherence* calls for sharing evidence among points with compatible structural features, *metric fidelity* requires the current local surface arrangement, and *persistent identity* provides a stable reference as visibility changes. These requirements motivate adaptive superpoints, geometric anchors, and canonical memory, respectively. As illustrated in Figure 2 (b), all three branches receive (P_t, F_t^{base}) and return features aligned with the current points. Below, $f_i \in F_{t,i}^{base}$, brackets denote concatenation, and ϕ and ψ denote learned MLPs.

Semantic branch. Spatial proximity alone can mix unrelated garment regions during self-contact. We aggregate regional evidence using both coordinates and features. FPS initializes representatives,

nearest-center assignment forms groups, and group-wise averaging produces superpoint coordinates s_j and features e_j. For point p_i, we retrieve a spatially local candidate set $\mathcal{N}_s(i)$. Separate MLPs encode the geometric relation $[p_i, s_j, p_i - s_j, \|p_i - s_j\|_2]$ and feature relation $[f_i, e_j, f_i - e_j]$ into r_{ij}^p and r_{ij}^f. With point queries $u_i^p = \psi_p(p_i)$ and $u_i^f = \psi_f(f_i)$, the association is

$$\alpha_{ij} = \mathrm{softmax}_{j \in \mathcal{N}_s(i)} \left(\frac{\langle r_{ij}^p, u_i^p \rangle + \langle r_{ij}^f, u_i^f \rangle}{\sqrt{h}} \right), \tag{1}$$

where h is the embedding dimension. One refinement step updates each feature by $\widetilde{e}_j = \sum_i \alpha_{ij} f_i / \max(\sum_i \alpha_{ij}, \epsilon_s)$, with the same weighted mean applied to coordinates; weights outside the candidate sets are zero. The structural context is returned to each point as $F_{t,i}^{sem} = \phi_s([f_i, \sum_{j \in \mathcal{N}_s(i)} \alpha_{ij} \widetilde{e}_j, p_i])$. This *dense-sparse-dense* exchange supplies regional evidence while retaining the original point alignment.

Geometry branch. Regional pooling can obscure the local arrangement of folds and contacting layers. To retain this information, we use spatial sampling and neighborhood aggregation. FPS selects anchors g_j from P_t, and $\mathcal{N}_g(j)$ contains their nearest observed points. The anchor representation explicitly encodes relative displacement and distance,

$$z_j^{geo} = \max_{i \in \mathcal{N}_g(j)} \phi_{local}\big([f_i, p_i - g_j, \|p_i - g_j\|_2]\big). \tag{2}$$

Channel-wise max pooling summarizes each neighborhood. Inverse distance interpolation returns the anchor features to the dense points, giving $\bar{z}_i^{geo} = \mathrm{Interp}_3(p_i; \{g_j, z_j^{geo}\})$ and $F_{t,i}^{geo} = \phi_g([f_i, \bar{z}_i^{geo}, p_i])$. Here, Interp_k denotes normalized inverse-distance interpolation over k nearest source points. The branch thus complements learned regional grouping with explicit current-frame metric relations.

Consistency branch. Current-frame structure and geometry provide limited evidence about regions that disappear and reappear. We retain a canonical reference in $C_{t-1} = \{(q_j, x_j^{t-1}, \Delta x_j^{t-1}, f_j^{t-1})\}_{j=1}^{A}$. Each anchor stores its fixed canonical coordinate, previous task-space position, recent displacement, and cached feature. We encode these quantities as $z_j = \phi_k([q_j, x_j^{t-1}, \Delta x_j^{t-1}, f_j^{t-1}])$ and construct the current query $u_i = \phi_q([f_i, p_i])$. The nearest cached positions define $\mathcal{N}_c(i)$, over which memory is retrieved by

$$c_i^{temp} = \sum_{j \in \mathcal{N}_c(i)} \frac{\exp(u_i^\top z_j / \sqrt{h})}{\sum_{\ell \in \mathcal{N}_c(i)} \exp(u_i^\top z_\ell / \sqrt{h})} z_j. \tag{3}$$

The output $F_{t,i}^{con} = \phi_c([f_i, c_i^{temp}, p_i])$ combines this canonical-indexed history with the current observation.

Adaptive gated residual fusion. A joint predictor applies a sigmoid to the concatenated base and branch features to obtain point-wise scalar gates g_i^s, g_i^g, g_i^c. Their weighted features enter a fusion MLP to obtain $\widetilde{F}_{t,i}^{base} = f_i + \lambda_{res}\, \phi_{fuse}\big([f_i, g_i^s F_{t,i}^{sem}, g_i^g F_{t,i}^{geo}, g_i^c F_{t,i}^{con}]\big)$, where $\lambda_{res} = \sigma(\gamma)$ is a learned scalar shared across points and initialized to give a small residual contribution. The shared prediction objectives below train the three representations through this enhanced feature.

3.4 Refinement and Reconstruction

Canonical refinement. We follow the NOCS refinement and volumetric reconstruction pipeline of GarmentTracking (Xue et al., 2023), replacing its fused point features with $\widetilde{F}_t^{base}$. The NOCS head predicts logits L_t^{raw} over B bins per axis, with bin coordinates $c_k = k/(B-1)$. Axis-wise argmax yields $\widehat{Y}_t^{raw}$. A PointNet encoder processes the raw NOCS, task-space coordinates, logits, and enhanced features; its pooled context is combined with PointNet features of Q_t^{raw} to form a joint descriptor. The point head combines local point features with this descriptor and adds residual logits, $L_t^{ref} = \mathrm{sg}(L_t^{raw}) + \Delta L_t$, while the mesh head predicts a shared scale s_t and offset b_t,

$$\widehat{q}_{tm} = (q_{tm}^{raw} - o) \odot s_t + o + b_t, \qquad o = (1/2, 1/2, 1/2). \tag{4}$$

Here, sg is stop gradient, o is the fixed center of the unit NOCS cube and $\odot$ denotes element-wise multiplication. The refiner receives detached input logits and point features, and the mesh-to-point

global context is also detached. Refined logits yield the hard point NOCS $\widehat{Y}_t$ used for volumetric aggregation.

Canonical supervision. We write losses for one training pair and average over the mini-batch, except for the explicitly batch-weighted term below. A superscript $*$ denotes ground truth, N and M are the observed-point and surface-sample counts. Let π^r_{iak} be the predicted bin probability for point i, axis a, and stage $r \in \{raw, ref\}$, and let k^*_{ia} be its ground-truth bin. Both stages use $\mathcal{L}^r_{ce} = -\frac{1}{3N}\sum_{i=1}^{N}\sum_{a=1}^{3}\log \pi^r_{iak^*_{ia}}$.

To preserve continuous coordinate distances, we also compute $\widehat{y}^{soft}_{ia} = \sum_{k=0}^{B-1} c_k \pi^{ref}_{iak}$ and minimize $\mathcal{L}_{coord} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\sum_{a=1}^{3} w_a(\widehat{y}^{soft}_{ia} - y^*_{ia})^2 + \epsilon}$, $(w_1, w_2, w_3) = (1, 1, w_z)$. The axis weight w_z controls the contribution of the third canonical coordinate. Ordinal supervision assigns nearby bins similar targets. Specifically, $g_{iak} = v_{iak}/(\sum_\ell v_{ia\ell} + \epsilon)$, where $v_{iak} = \exp[-(k - (B-1)y^*_{ia})^2/(2\sigma^2_{ord})]$ for ground-truth coordinates clipped to $[0, 1]$. With bandwidth σ_{ord}, the objective is

$$\mathcal{L}_{ord} = \frac{1}{N}\sum_{i=1}^{N}\sum_{a=1}^{3}\sum_{k=0}^{B-1} g_{iak}\log\frac{g_{iak}}{\pi^{ref}_{iak}}. \tag{5}$$

Small positive constants ϵ stabilize the computations. The refined canonical surface receives paired coordinate supervision, $\mathcal{L}_{mesh} = \|\widehat{Q}_t - Q^*_t\|^2_F/(3M)$, which trains its predicted scale and offset.

Volumetric reconstruction. The canonical volumetric encoder transforms $[\widetilde{F}^{base}_{t,i}, \widehat{y}_{t,i} - v_i, p_i]$ with an MLP, where v_i is the grid point assigned to $\widehat{y}_{t,i}$. Features are scattered by max reduction onto a 32^3 grid and processed by a four-level 3D U-Net to obtain V_t. For each refined canonical query, the decoder samples the volume by trilinear interpolation and predicts its task-space position,

$$\widehat{x}_{tm} = W_t(\widehat{q}_{tm}; V_t) = \phi_W(\mathrm{Tri}(V_t, 2\widehat{q}_{tm} - 1)). \tag{6}$$

The normalization maps queries to the sampler's coordinate range. The resulting $\widehat{X}_t = \{\widehat{x}_{tm}\}_{m=1}^{M}$ is the reconstructed garment surface, and the same decoder is queried for memory updates.

Surface supervision. For matched surface samples, let $d_m = \widehat{x}_{tm} - x^*_{tm}$. We combine coordinate MSE, $\mathcal{L}_{warp} = (3M)^{-1}\sum_m \|d_m\|_2^2$, with Euclidean correspondence error, $\mathcal{L}_{corr} = M^{-1}\sum_m \|d_m\|_2$. To supervise surface coverage, we select the same deterministic, evenly spaced index subset from prediction and ground truth, obtaining point sets $\widehat{\mathcal{S}}$ and $\mathcal{S}^*$ of size R. The symmetric surface objective is

$$\mathcal{L}_{surf} = \frac{1}{2R}\Big(\sum_{\widehat{x}\in\widehat{\mathcal{S}}}\min_{x^*\in\mathcal{S}^*}\|\widehat{x} - x^*\|_2 + \sum_{x^*\in\mathcal{S}^*}\min_{\widehat{x}\in\widehat{\mathcal{S}}}\|x^* - \widehat{x}\|_2\Big). \tag{7}$$

Local edge vectors additionally preserve the target surface arrangement. For R_g sampled query indices $\mathcal{I}_g$, let $\mathcal{N}^*(m)$ contain K_g nearest neighbors in the full ground-truth surface, excluding the query itself. We minimize $\mathcal{L}_{geo} = \frac{1}{R_g K_g}\sum_{m\in\mathcal{I}_g}\sum_{r\in\mathcal{N}^*(m)} H_{\beta_g}[(\widehat{x}_{tr} - \widehat{x}_{tm}) - (x^*_{tr} - x^*_{tm})]$, where $H_\beta(v) = \frac{1}{3}\sum_{a=1}^{3}\rho_\beta(v_a)$, with $\rho_\beta(d) = d^2/(2\beta)$ for $|d| < \beta$ and $\rho_\beta(d) = |d| - \beta/2$ otherwise. β is the Smooth-L_1 transition parameter.

To emphasize training pairs near the correspondence-accuracy threshold, let $e_b = M^{-1}\sum_m \|d_{bm}\|_2$ for sample b in a batch of size n. Detached weights $\omega_b = \exp[-(\mathrm{sg}(e_b) - \mu)^2/(2\sigma^2_{bd})]$ focus optimization around error μ with bandwidth σ_{bd}. The threshold-aware objective is

$$\mathcal{L}_{bound} = \frac{\sum_{b=1}^{n}\omega_b \max(0, e_b - m_{target})}{\sum_{b=1}^{n}\omega_b + \epsilon_b}, \tag{8}$$

where m_{target} is the target error margin and $\epsilon_b > 0$ stabilizes normalization.

3.5 Canonical Memory Update

FPS initializes A canonical anchors from the previous canonical surface. Their positions are interpolated from its corresponding task-space samples, and their features from the previous point NOCS

Table 1: **Quantitative comparison on VR-Folding dataset.** Results are reported for four garment categories under both folding and flattening tasks. Higher A_d and lower D_{corr}, D_{chamf}, and D_{nocs} indicate better performance.

Type	Method	Folding					Flattening				
		A_{3cm} ↑	A_{5cm} ↑	D_{corr} ↓	D_{chamf} ↓	D_{nocs} ↓	A_{5cm} ↑	A_{10cm} ↑	D_{corr} ↓	D_{chamf} ↓	D_{nocs} ↓
Shirt	GarmentNets Chi & Song (2021)	0.8%	21.5%	6.40	1.58	0.221	13.2%	59.4%	10.54	3.54	0.135
	GarmentTracking Xue et al. (2023)	29.8%	85.8%	3.88	1.16	0.051	30.7%	83.4%	8.63	1.75	0.105
	GaPT-DAR Zhang et al. (2025)	32.3%	89.3%	3.85	1.03	0.044	33.1%	87.6%	8.61	1.68	0.098
	SCLTrack (Ours)	**40.1%**	**90.5%**	**3.66**	**1.01**	**0.043**	**42.7%**	**89.1%**	**8.50**	**1.59**	**0.091**
Pants	GarmentNets Chi & Song (2021)	16.2%	69.5%	4.43	1.30	0.162	1.5%	42.4%	12.54	4.19	0.185
	GarmentTracking Xue et al. (2023)	47.3%	94.0%	3.26	1.07	0.039	31.3%	78.2%	8.97	1.64	0.113
	GaPT-DAR Zhang et al. (2025)	49.1%	95.2%	3.15	0.99	0.031	32.8%	79.3%	8.79	1.44	0.101
	SCLTrack (Ours)	**53.4%**	**96.9%**	**3.04**	**0.93**	**0.022**	**34.4%**	**82.0%**	**8.61**	**1.27**	**0.092**
Top	GarmentNets Chi & Song (2021)	10.3%	53.8%	5.19	1.51	0.148	21.6%	57.6%	9.98	2.13	0.174
	GarmentTracking Xue et al. (2023)	37.9%	85.9%	3.75	0.99	0.051	36.5%	69.0%	9.41	1.59	0.113
	GaPT-DAR Zhang et al. (2025)	39.5%	88.3%	3.59	0.89	0.045	37.6%	73.2%	9.38	1.48	0.107
	SCLTrack (Ours)	**45.2%**	**89.8%**	**3.42**	**0.77**	**0.036**	**39.5%**	**76.7%**	**9.22**	**1.39**	**0.101**
Skirt	GarmentNets Chi & Song (2021)	1.1%	30.3%	6.95	1.89	0.239	0.1%	7.9%	18.48	5.99	0.287
	GarmentTracking Xue et al. (2023)	23.8%	71.3%	4.61	1.33	0.060	5.4%	39.4%	16.09	2.02	0.199
	GaPT-DAR Zhang et al. (2025)	25.1%	73.1%	4.55	1.25	0.048	8.0%	39.9%	15.98	1.99	0.187
	SCLTrack (Ours)	**28.4%**	**75.6%**	**4.43**	**1.19**	**0.038**	**10.3%**	**40.8%**	**15.82**	**1.90**	**0.179**

Table 2: **Ablation of structured correspondence components.** We evaluate the complementary contributions of the semantic (S), geometry (G), and consistency (C) branches, as well as adaptive gated residual fusion (AGF). A_3 and A_{10} respectively correspond to the results of Folding and Flattening while the rest are averages of both.

S	G	C	AGF	A_{3cm} ↑	A_{5cm} ↑	A_{10cm} ↑	D_{corr} ↓	D_{chamf} ↓	D_{nocs} ↓
				34.7%	55.1%	67.5%	7.33	1.44	0.091
✓				35.8%	56.4%	68.6%	7.28	1.37	0.085
	✓			36.1%	56.2%	68.9%	7.23	1.32	0.088
		✓		37.5%	58.0%	70.8%	7.19	1.32	0.079
✓	✓		✓	37.8%	57.8%	71.1%	7.18	1.31	0.078
✓		✓	✓	39.9%	58.4%	71.4%	7.14	1.32	0.077
	✓	✓	✓	40.5%	58.5%	71.5%	7.12	1.29	0.079
✓	✓	✓		41.2%	59.6%	71.9%	7.10	1.28	0.076
✓	✓	✓	✓	**41.8%**	**60.0%**	**72.2%**	**7.09**	**1.26**	**0.075**

and projected encoder features; initial displacements are zero. After reconstruction, each identity q_j is retained and queried through the shared warp to obtain $x_j^t = W_t(q_j; V_t)$. We refresh its feature by $f_j^t = \mathrm{Interp}_3(x_j^t; P_t, \widetilde{F}_t^{base})$ and store $C_t = \left\{\left(q_j, x_j^t, x_j^t - x_j^{t-1}, f_j^t\right)\right\}_{j=1}^{A}$. The displacement is $\Delta x_j^t = x_j^t - x_j^{t-1}$. Training constructs a fresh previous-frame cache for each pair, whereas online inference passes the detached updated cache to the next frame.

Anchor supervision. We supervise the shared warp at these persistent identities using targets $\bar{x}_j^t = \mathrm{Interp}_3(q_j; Q_t^*, X_t^*)$. The anchor objective is

$$\mathcal{L}_{anchor} = \frac{\lambda_c(e)}{A} \sum_{j=1}^{A} H_{\beta_c}(x_j^t - \bar{x}_j^t), \tag{9}$$

where β_c controls the robust penalty and $\lambda_c(e) = \lambda_c^* \,\mathrm{clip}((e - e_0)/E_w, 0, 1)$ increases to its target weight over $E_w > 0$ epochs after epoch e_0. This supervision aligns canonical identities with their current task-space locations. The final total loss function is defined as the weighted sum of all the losses above.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate on VR-Folding (Xue et al., 2023) dataset. Following GarmentTracking, D_{nocs} is the mean Euclidean NOCS error, D_{corr} measures task-space correspondence error, and D_{chamf} is the symmetric surface distance; the latter two are reported in centimeters. A_d denotes the percentage of frames with $D_{corr} < d$ cm. We report A_3/A_5 for folding and A_5/A_{10} for flattening. Following prior work, we use the stricter A_3/A_5 thresholds for folding and A_5/A_{10}

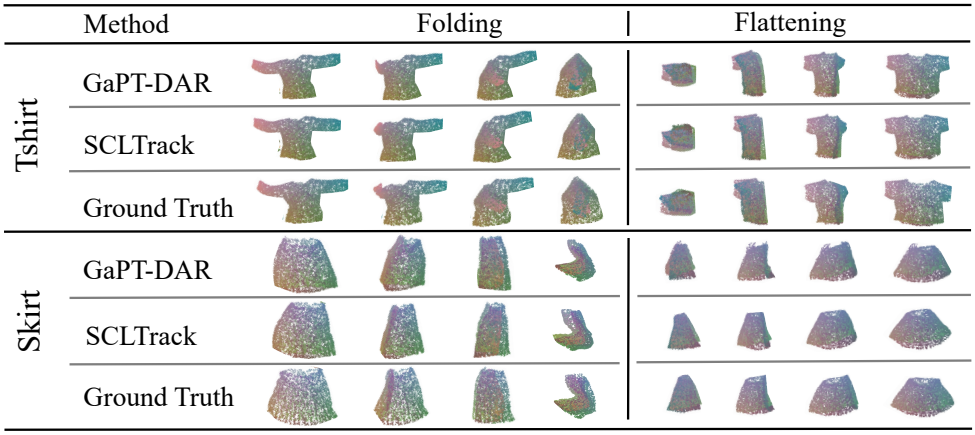

Figure 3: **Qualitative comparison of category-level garment pose tracking.** GaPT-DAR and SCLTrack are compared with ground truth on T-shirt and skirt sequences under folding and flattening, with four temporal points shown for each task.

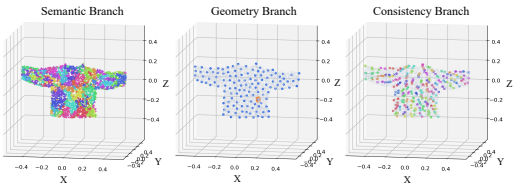

Figure 4: **Visualization of structured branch representations.** The semantic, geometry, and consistency branches capture complementary regional, metric, and canonical information.

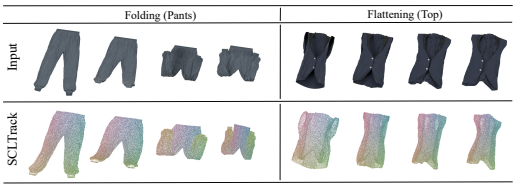

Figure 5: **Qualitative results on real-world garment sequences.** SCLTrack preserves coherent garment reconstruction across folding and flattening manipulations.

for flattening to reflect the different correspondence error scales induced by their deformation and motion ranges. For ablations, categories are averaged equally, with A_3 computed on folding, A_{10} on flattening, and the remaining metrics over all eight category-task combinations.

Implementation details. Each training pair uses 4,000 points per observation, voxelized at 2.5 mm, and 6,000 surface queries. The fine-tuning configuration uses Adam for 12 epochs with batch size 8, a base learning rate of 5×10^{-6}, and gradient-norm clipping at 1.0. NOCS and backbone learning-rate multipliers are 0.5 and 0.01. Training builds a previous-frame cache for each pair, whereas inference updates and propagates the cache sequentially. All experiments are conducted on NVIDIA RTX PRO 6000 GPUs.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares SCLTrack with GarmentNets (Chi & Song, 2021), GarmentTracking (Xue et al., 2023), and GaPT-DAR (Zhang et al., 2025). SCLTrack achieves the best values for all reported metrics across the four categories and both tasks. Relative to GaPT-DAR, equal-weight averaging over the eight category-task combinations gives reductions of 2.1%, 6.5%, and 8.9% in D_{corr}, D_{chamf}, and D_{nocs}, respectively. Joint improvement in canonical coordinates, sensitive identity correspondence, and surface proximity indicates gains beyond geometric overlap alone. On folding, mean A_3 rises from 36.5% to 41.8%, while mean A_5 increases from 86.5% to 88.2%, showing a larger gain at the stricter accuracy threshold. Bewsides, the category-level results further distinguish these effects. For folding Pants, D_{nocs} decreases from 0.031 to 0.022; for folding Top, D_{chamf} decreases from 0.89 to 0.77. On flattening Shirt, A_5 improves by 9.6 percentage points despite a smaller change in mean correspondence distance, indicating a shift in the proportion of accurately tracked frames. Skirt flattening remains challenging, with $D_{corr} = 15.82$ and $D_{chamf} = 1.90$, emphasizing the dis-

tinction between recovering a plausible surface and assigning correct correspondences. Qualitative results are shown in Figure 3, and the real-world tracking results are shown in Figure 5.

Tracking Speed. SCLTrack achieves an inference speed of 21 FPS when running on a single NVIDIA RTX PRO 6000 GPU, which is competitive with existing methods for the same task.

Table 3: **Spatial representation ablation.** S/G denotes the modified semantic/geometry branch, while all other components are kept fixed.

Variant	$A_{5\text{cm}}\uparrow$	$D_{\text{corr}}\downarrow$	$D_{\text{chamf}}\downarrow$	$D_{\text{nocs}}\downarrow$
S: pointwise	59.2%	7.13	1.30	0.080
S: hard groups	59.4%	7.10	1.28	0.079
S: coord. only	59.5%	7.11	1.27	0.081
G: random anchors	59.8%	7.11	1.29	0.077
G: w/o rel. geom.	59.9%	7.12	1.29	0.078
Full SCLTrack	**60.0%**	**7.09**	**1.26**	**0.075**

Table 4: **Consistency branch ablation.** Canonical memory components are modified independently, † means inference-only interventions.

Variant	$A_5\uparrow$	$D_{\text{corr}}\downarrow$	$D_{\text{chamf}}\downarrow$	$D_{\text{nocs}}\downarrow$
w/o q_j in token	58.9%	7.14	1.27	0.079
Uniform attention	59.0%	7.15	1.29	0.079
Frozen positions†	59.4%	7.14	1.29	0.081
Frozen features†	59.3%	7.12	1.30	0.080
w/o $\mathcal{L}_{anchor}$	59.8%	7.10	1.28	0.077
Full SCLTrack	**60.0%**	**7.09**	**1.26**	**0.075**

4.3 Ablation Studies

Branch composition and fusion. Table 2 evaluates semantic (S), geometry (G), consistency (C), and adaptive gated residual fusion (AGF). S obtains lower D_{nocs} than G, 0.085 versus 0.088, whereas G gives lower D_{chamf}, 1.32 versus 1.37. C reaches $A_5 = 58.0\%$, compared with 55.1% for the baseline. Among the two-branch configurations, all using AGF, S+C provides the lowest NOCS error, while G+C gives the lowest correspondence and surface distances. These tradeoffs support combining regional, metric, and historical evidence. With all branches present, AGF predicts point-wise branch weights and improves A_5 from 59.6% to 60.0%, while reducing D_{chamf} from 1.28 to 1.26. The full model therefore benefits from both complementary inputs and their adaptive integration.

Spatial representation. Table 3 modifies one spatial mechanism at a time. For the semantic branch, *pointwise* replaces superpoint aggregation with a point-wise transformation. *Hard groups* retains the initial nearest-center partition without soft refinement, while *coord. only* removes the feature relation from association scores. Their NOCS errors of 0.080, 0.079, and 0.081 exceed the full model's 0.075. The coordinate variant alone further supports feature-aware regional association. For geometry, *random anchors* replaces FPS with equal-count random sampling, and *w/o rel. geom.* omits explicit relative displacement and distance while retaining neighborhood aggregation. Both increase D_{chamf} to 1.29, linking spatial coverage and metric encoding to reconstruction quality. Figure 4 illustrates the distinct regional and anchor-based representations underlying these comparisons.

Consistency branch and canonical memory. Table 4 evaluates the consistency branch by separately analyzing canonical identity, memory retrieval, state updates, and anchor supervision. Removing q_j from the token leaves the anchors in place but removes their explicit canonical-coordinate input, and uniform attention equally weights the same local candidate anchors. These changes reduce A_5 to 58.9% and 59.0%, respectively, supporting identity-aware and selective memory retrieval. Frozen positions and frozen features retain the cached spatial states or feature values during inference, without retraining. They reduce A_5 to 59.4% and 59.3%, showing that refreshing both types of state contributes to tracking. Finally, removing $\mathcal{L}_{anchor}$ retains the memory mechanism but disables its auxiliary warp supervision, giving $A_5 = 59.8\%$ and $D_{chamf} = 1.28$. These results suggest that canonical inputs, selective retrieval, and state updates contribute to the observed performance.

5 Conclusion

In this work, we presented SCLTrack, a structured correspondence learning framework for category-level garment pose tracking. SCLTrack organizes correspondence reasoning into complementary semantic, geometric, and identity clues through adaptive superpoints, FPS-based local aggregation, and recurrent canonical memory. These representations are fused and decoded through canonical refinement and volumetric warping for garment reconstruction. Experiments on VR-Folding show consistent improvements in correspondence and reconstruction, while ablations provide evidence for the complementary contributions of the three branches and canonical memory. Real-world qualitative results further demonstrate the applicability of SCLTrack to captured garment observations.

AI USE STATEMENT

In this study, we used generative AI tools to edit text written by the authors, with a view to improving its clarity, grammatical accuracy and readability. We did not use generative AI tools for research methodology, experimental design, methodological implementation, data analysis, interpretation of results or the formulation of scientific conclusions. All AI-assisted edits were reviewed by the authors. We take full responsibility for the final content of this study.

REPRODUCIBILITY STATEMENT

The main paper specifies the SCLTrack architecture, training objectives, implementation details, and evaluation protocol. **Our training, prediction and evaluation code, together with the corresponding configuration files, are released as anonymous supplementary material to support reproduction of the reported experiments.** The specific implementation details are explained in the README file in the supplementary materials.

REFERENCES

Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15854–15864. IEEE, 2021.

Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2753–2762. IEEE, 2021.

Zhangquan Chen, Puhua Jiang, and Ruqi Huang. Dv-matcher: Deformation-based non-rigid point cloud matching guided by pre-trained visual features. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27264–27274. IEEE, 2025.

Cheng Chi and Shuran Song. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3324–3333, October 2021.

Ronghua Du, Zong Li, Jinlai Zhang, Kai Gao, and Lin Hu. Point cloud mapping and loop closure detection using superpoint semantic graph for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2026.

Louis Geist, Loic Landrieu, and Damien Robert. Ez-sp: Fast and lightweight superpoint-based 3d segmentation. *arXiv preprint arXiv:2512.00385*, 2025.

Rhodri Guerrier, Adam W Harley, and Dima Damen. Pointst3r: Point tracking through 3d grounded correspondence. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7430–7439. IEEE, 2026.

Le Hui, Linghua Tang, Yaqi Shen, Jin Xie, and Jian Yang. Learning superpoint graph cut for 3d instance segmentation. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. doi: 10.52202/068431-2667.

Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5606–5616. IEEE, 2022.

Yair Kittenplon, Yonina C Eldar, and Dan Raviv. Flowstep3d: Model unrolling for self-supervised scene flow estimation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4112–4121. IEEE, 2021.

Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8490–8499. IEEE, 2022.

Itai Lang, Dror Aiger, Forrester Cole, Shai Avidan, and Michael Rubinstein. Scoop: Self-supervised correspondence and optimization-based scene flow. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5281–5290. IEEE, 2023.

Ren Li, Corentin Dumery, Zhantao Deng, and Pascal Fua. Reconstruction of manipulated garment with guided deformation prior. *Advances in Neural Information Processing Systems*, 37:58637–58662, 2024.

Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3706–3715, 2020.

Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5544–5554. IEEE, 2022.

Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3540–3549. IEEE, 2021.

Liu Liu, Anran Huang, Qi Wu, Dan Guo, Xun Yang, and Meng Wang. KPA-Tracker: Towards robust and real-time category-level articulated object 6d pose tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 3684–3692, 2024. doi: 10.1609/aaai.v38i4.28158.

Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv preprint arXiv:2202.07123*, 2022.

David Novotny, Ignacio Rocco, Samarth Sinha, Alexandre Carlier, Gael Kerchenbaum, Roman Shapovalov, Nikita Smetanin, Natalia Neverova, Benjamin Graham, and Andrea Vedaldi. Keytr: Keypoint transporter for 3d reconstruction of deformable objects in videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5585–5594. IEEE, 2022.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5099–5108. Curran Associates, Inc., 2017.

Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022. doi: 10.52202/068431-1685.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11143–11152, 2022.

Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2393–2401, 2023.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2637–2646. IEEE, 2019.

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 606–617. IEEE, 2023.

Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13189–13198. IEEE, 2021.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in neural information processing systems*, 35:33330–33342, 2022.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851. IEEE, 2024.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: Advancing 3d point tracking with explicit camera motion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6726–6737. IEEE, 2025.

Han Xue, Wenqiang Xu, Jieyi Zhang, Tutian Tang, Yutong Li, Wenxin Du, Ruolin Ye, and Cewu Lu. Garmenttracking: Category-level garment pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21233–21242, June 2023.

Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, and Slobodan Ilic. Rotation-invariant transformer for point cloud matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5384–5393. IEEE, 2023.

Qiaojun Yu, Ce Hao, Xibin Yuan, Li Zhang, Liu Liu, Yukang Huo, Rohit Agarwal, and Cewu Lu. Generalizable articulated object perception with superpoints. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10890874.

Li Zhang, Mingliang Xu, Jianan Wang, Qiaojun Yu, Lixin Yang, Yonglu Li, Cewu Lu, Rujing Wang, and Liu Liu. GaPT-DAR: Category-level garments pose tracking via integrated 2d deformation and 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22638–22647, June 2025.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

Xingyu Zhu, Yan Wu, Zhiwen Tu, Haifeng Zhong, and Yixing Gao. Generalizable category-level topological structure learning for clothing recognition in robotic grasping. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 21259–21266. IEEE, 2025.

APPENDIX

A OVERVIEW

This supplementary material provides additional details for the task formulation and experimental evaluation of SCLTrack. Section B introduces the background of category-level garment pose tracking, including Normalized Object Coordinate Space (NOCS), garment pose representation, and the definition of sequential garment pose tracking. Section C extends the experimental description in the main paper. We first introduce the VR-Folding benchmark, including its VR-based data collection procedure, task protocols, dataset statistics, and data split. We then provide the mathematical definitions of the six evaluation metrics used in our experiments. Finally, we present extended qualitative results for all four garment categories under both folding and flattening manipulations. Additionally, we also provide details regarding the reproducibility of this paper in the Reproducibility Statement to ensure that the results can be reproduced.

B PRELIMINARIES

B.1 NORMALIZED OBJECT COORDINATE SPACE

Normalized Object Coordinate Space (NOCS) provides a category-level canonical coordinate system in which object instances with different shapes can share a common spatial reference (Wang et al., 2019). Following GarmentNets (Chi & Song, 2021) and GarmentTracking (Xue et al., 2023), we use NOCS to represent the canonical state of garments. For each garment instance, the rest-state garment mesh is represented in a normalized canonical space, with the garment worn in the canonical T-pose provided by CLOTH3D.

Let the canonical garment surface be

$$Q = \{q_m\}_{m=1}^{M}, \qquad q_m \in [0,1]^3, \tag{10}$$

where q_m denotes the normalized canonical coordinate of surface point m. At frame t, the same material point may move to a task-space position $x_{tm} \in \mathbb{R}^3$ due to non-rigid deformation, while its canonical coordinate q_m remains unchanged. Hence, the correspondence between canonical and task space can be written as

$$q_m \longleftrightarrow x_{tm}. \tag{11}$$

For a partial point-cloud observation

$$P_t = \{p_{ti}\}_{i=1}^{N_t}, \tag{12}$$

each visible point p_{ti} is associated with a NOCS coordinate $y_{ti} \in [0,1]^3$. The NOCS representation therefore serves as a deformation-independent identity of an observed garment point. Points that undergo large displacement, folding, or self-contact in task space can still be associated through their canonical coordinates.

This property is particularly useful for garments because their task-space geometry varies substantially over time, whereas the canonical coordinate of each material location provides a stable correspondence reference.

B.2 GARMENT POSE

Object pose describes the configuration of an object in the observation space. For a rigid object, its pose can typically be parameterized by a 6-DoF rigid transformation consisting of three rotational and three translational degrees of freedom. Garments, in contrast, are highly deformable and possess near-infinite degrees of freedom (Xue et al., 2023). Their pose is therefore represented by the spatial configuration of the complete garment surface rather than by a single rigid transformation.

Given the canonical surface Q, we denote the non-rigid canonical-to-task mapping at frame t by

$$W_t : \mathbb{R}^3 \to \mathbb{R}^3. \tag{13}$$

The complete garment pose in task space is

$$X_t = \{x_{tm}\}_{m=1}^{M}, \qquad x_{tm} = W_t(q_m). \tag{14}$$

Thus, a garment pose can equivalently be represented by the set of paired canonical and task-space coordinates

$$\mathcal{G}_t = \{(q_m, x_{tm})\}_{m=1}^{M}. \tag{15}$$

The canonical coordinate identifies a material location on the garment, while the task-space coordinate specifies where that location is positioned under the current deformation. This representation describes both the complete garment geometry and the dense correspondence between canonical and observed spaces.

B.3 Category-Level Garment Pose Tracking

Category-level garment pose tracking extends garment pose estimation from an individual observation to a sequence of partial observations. Given a point-cloud sequence

$$\mathcal{P}_{1:T} = \{P_1, P_2, \ldots, P_T\}, \tag{16}$$

the goal is to recover the garment configuration at every frame while preserving the canonical identity of corresponding material points over time.

For each frame t, the tracker estimates the per-point canonical correspondence

$$\hat{Y}_t = \{\hat{y}_{ti}\}_{i=1}^{N_t}, \tag{17}$$

together with the complete task-space garment surface

$$\hat{X}_t = \{\hat{x}_{tm}\}_{m=1}^{M}. \tag{18}$$

The sequential prediction objective can therefore be expressed as

$$\mathcal{P}_{1:T} \longrightarrow \left\{(\hat{Y}_t, \hat{X}_t)\right\}_{t=1}^{T}. \tag{19}$$

In an online setting, the prediction at frame t uses the current observation together with information propagated from previous frames. In SCLTrack, this is instantiated as

$$(\hat{Y}_t, \hat{X}_t) = \mathcal{F}\left(P_{t-1}, P_t, Y_{t-1}, Q_t^{\mathrm{raw}}; C_{t-1}\right), \tag{20}$$

where C_{t-1} denotes the recurrent canonical memory. The persistent canonical coordinates provide the identity reference, while the task-space surface evolves continuously with garment deformation. Category-level tracking further requires the learned model to generalize this representation to previously unseen garment instances from the same semantic category.

C Experiments, Extended

C.1 VR-Folding Dataset

Dataset overview. We conduct experiments on the VR-Folding dataset introduced by GarmentTracking (Xue et al., 2023). The garment assets are derived from CLOTH3D and cover four categories: *Shirt*, *Pants*, *Top*, and *Skirt*. VR-Folding contains two manipulation tasks, *Flattening* and *Folding*, which generate substantially different non-rigid garment configurations. In total, the dataset contains 9,767 manipulation videos and approximately 790K multi-view RGB-D frames with complete garment pose and hand-pose annotations.

VR-based data collection. VR-Folding is collected using the VR-Garment recording system. A volunteer wears an HTC Vive Pro headset and VR gloves, which transfer real-world hand motion to virtual hands in a Unity simulation environment. Cloth dynamics are simulated in Unity, and the interaction system allows the virtual fingers to grasp or release arbitrary points on the garment surface.

At the beginning of each recording, a garment instance is placed in the virtual scene. The volunteer performs the predefined manipulation while the simulator records the complete deformed garment mesh and hand poses at every simulation step. After an interaction sequence is completed, the recorded animation is re-rendered from multiple viewpoints. The re-rendering process produces

Table 5: Detailed statistics of the VR-Folding dataset. Instance counts are reported separately for each manipulation task.

Task	Statistic	Shirt	Pants	Top	Skirt	Total
Folding	Instances	989	1,538	903	461	3,891
	Videos	993	1,551	889	463	3,896
	Frames	66,830	77,479	37,214	23,427	204,950
Flattening	Instances	1,037	1,631	1,040	467	4,175
	Videos	2,145	1,720	1,063	943	5,871
	Frames	243,460	136,411	132,580	73,357	585,808

RGB images, depth images, garment masks, and complete garment meshes with NOCS annotations. Frames are rendered at 10 FPS, and each recorded frame contains four RGB-D views.

This collection strategy preserves natural human manipulation trajectories while retaining direct access to dense ground-truth garment geometry and canonical correspondence throughout the entire sequence.

Manipulation protocols. The *Flattening* task starts from a garment randomly dropped onto the virtual table. A surface point is selected, and the volunteer first lifts the garment before repeatedly grasping, swinging, and spreading it with both hands. The manipulation continues until the garment reaches a flattened T-pose.

The *Folding* task starts from the flattened configuration. The volunteer performs a sequence of bimanual pick-and-place operations following category-specific folding rules. Shirts are folded using two or three operations, depending on sleeve length. Pants, tops, and skirts are generally folded first along the left–right direction and subsequently along the up–down direction. Variations in grasping positions and manipulation trajectories yield diverse intermediate garment configurations.

Dataset statistics. Table 5 summarizes the detailed statistics. Each video contains one garment instance, while the same instance can appear in multiple manipulation videos.

The complete benchmark therefore contains 3,896 folding videos and 5,871 flattening videos, corresponding to 204,950 and 585,808 frames, respectively. The dataset is partitioned into training, validation, and test subsets with an 8:1:1 ratio by garment instance. Consequently, the garment instances used for validation and testing are distinct from the instances used for training. We use the same instance-level protocol with split seed zero in our experiments.

C.2 Evaluation Metrics

We follow the evaluation protocol of GarmentTracking (Xue et al., 2023). The six metrics used in our experiments are NOCS Coordinate Distance D_{nocs}, Correspondence Distance D_{corr}, Chamfer Distance D_{chamf}, and the three threshold accuracies A_3, A_5, and A_{10}.

Let the evaluation sequence contain T frames. At frame t, let

$$\hat{Y}_t = \{\hat{y}_{ti}\}_{i=1}^{N_t}, \qquad Y_t^* = \{y_{ti}^*\}_{i=1}^{N_t} \tag{21}$$

denote the predicted and ground-truth NOCS coordinates of the observed point cloud. We further denote the reconstructed task-space garment surface by

$$\hat{X}_t = \{\hat{x}_{tm}\}_{m=1}^{M_t}, \tag{22}$$

with associated canonical coordinates $\hat{Q}_t = \{\hat{q}_{tm}\}_{m=1}^{M_t}$, and the ground-truth surface by

$$X_t^* = \{x_{tn}^*\}_{n=1}^{M_t^*}, \qquad Q_t^* = \{q_{tn}^*\}_{n=1}^{M_t^*}. \tag{23}$$

NOCS Coordinate Distance. D_{nocs} evaluates the point-wise canonical correspondence of the partial observation. For each frame,

$$d_{\text{nocs}}^{(t)} = \frac{1}{N_t} \sum_{i=1}^{N_t} \|\hat{y}_{ti} - y_{ti}^*\|_2 \,. \tag{24}$$

The sequence-level metric is

$$D_{\text{nocs}} = \frac{1}{T} \sum_{t=1}^{T} d_{\text{nocs}}^{(t)}. \tag{25}$$

Since NOCS coordinates are normalized, D_{nocs} is unitless. A lower value indicates more accurate canonical correspondence.

Correspondence Distance. D_{corr} measures task-space error between corresponding garment surface points. Correspondences are established through canonical coordinates. For predicted surface point m, we first find its ground-truth counterpart by

$$c_t(m) = \arg\min_n \|\hat{q}_{tm} - q_{tn}^*\|_2 \,. \tag{26}$$

The mean correspondence error of frame t is then

$$d_{\text{corr}}^{(t)} = \frac{1}{M_t} \sum_{m=1}^{M_t} \left\| \hat{x}_{tm} - x_{t,c_t(m)}^* \right\|_2 . \tag{27}$$

After converting task-space distances to centimeters, the reported metric is

$$D_{\text{corr}} = \frac{1}{T} \sum_{t=1}^{T} d_{\text{corr}}^{(t)}. \tag{28}$$

Lower D_{corr} indicates more accurate identity-sensitive garment pose reconstruction.

Chamfer Distance. D_{chamf} evaluates geometric agreement between the reconstructed complete surface and the ground-truth garment surface, independently of explicit point identity. For frame t, the symmetric Chamfer distance is

$$\begin{aligned} d_{\text{chamf}}^{(t)} = \frac{1}{2} \Bigg[& \frac{1}{M_t} \sum_{m=1}^{M_t} \min_n \|\hat{x}_{tm} - x_{tn}^*\|_2 \\ & + \frac{1}{M_t^*} \sum_{n=1}^{M_t^*} \min_m \|x_{tn}^* - \hat{x}_{tm}\|_2 \Bigg]. \end{aligned} \tag{29}$$

The final metric is

$$D_{\text{chamf}} = \frac{1}{T} \sum_{t=1}^{T} d_{\text{chamf}}^{(t)}. \tag{30}$$

D_{chamf} is reported in centimeters, and a lower value indicates higher surface reconstruction accuracy.

Threshold Accuracy. The correspondence error can vary considerably across frames, particularly when the garment undergoes large deformation or self-contact. We therefore also use A_d, which measures the proportion of frames whose mean correspondence error falls below a distance threshold d:

$$A_d = \frac{100}{T} \sum_{t=1}^{T} \mathbb{I}\left[d_{\text{corr}}^{(t)} < d \right], \tag{31}$$

where $\mathbb{I}[\cdot]$ is the indicator function and d is measured in centimeters.

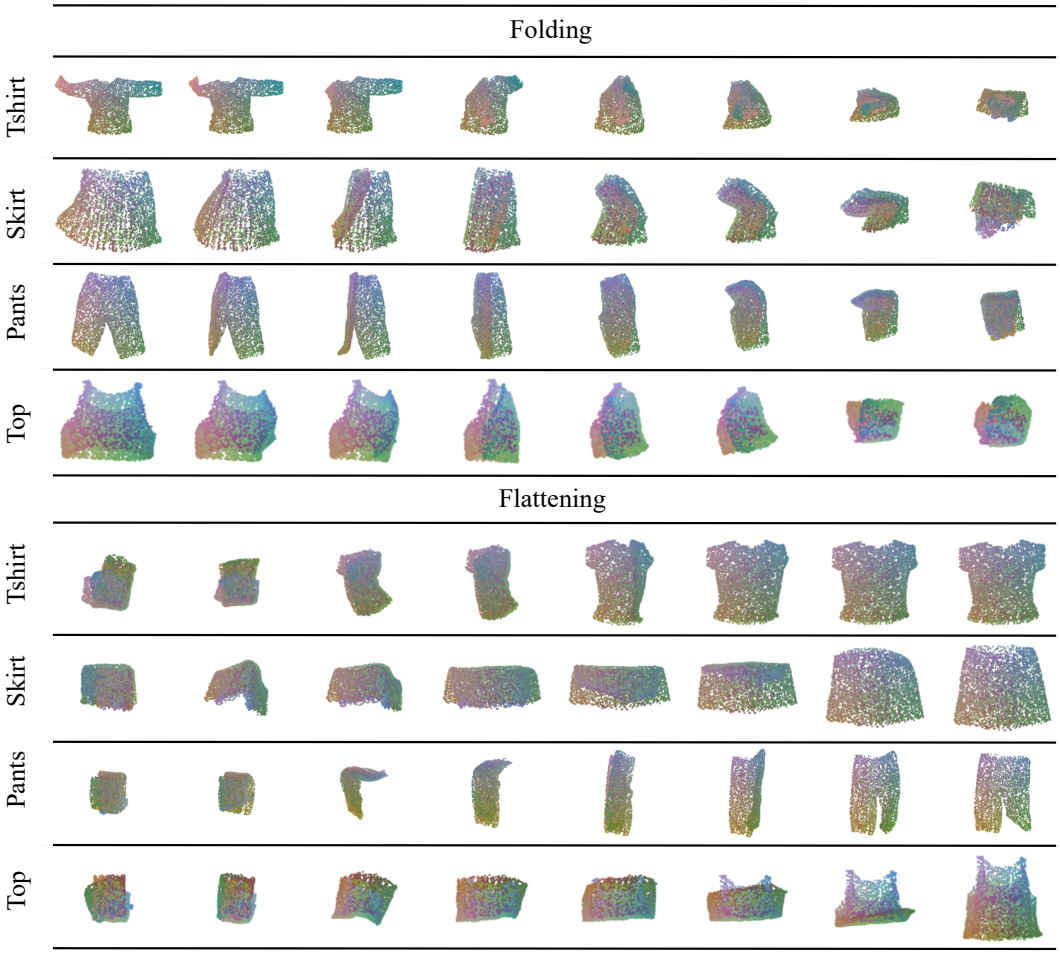

Figure 6: Supplementary qualitative results for four garment categories.

The three threshold metrics used in our experiments are therefore

$$A_3 = \frac{100}{T} \sum_{t=1}^{T} \mathbb{I}\left[d_{\text{corr}}^{(t)} < 3 \text{ cm}\right], \tag{32}$$

$$A_5 = \frac{100}{T} \sum_{t=1}^{T} \mathbb{I}\left[d_{\text{corr}}^{(t)} < 5 \text{ cm}\right], \tag{33}$$

$$A_{10} = \frac{100}{T} \sum_{t=1}^{T} \mathbb{I}\left[d_{\text{corr}}^{(t)} < 10 \text{ cm}\right]. \tag{34}$$

Higher A_d indicates that a larger fraction of the sequence is tracked within the specified correspondence tolerance. Following the established VR-Folding protocol, we report A_3 and A_5 for the folding task, and A_5 and A_{10} for the flattening task. The tighter thresholds used for folding and the wider thresholds used for flattening reflect the different correspondence error ranges induced by their manipulation and deformation patterns.

C.3 Qualitative Results, Extended

As shown in Figure 6, we provide additional qualitative results for all four garment categories, including Shirt, Pants, Top, and Skirt, under both folding and flattening manipulations. Compared with

the examples in the main paper, these results include longer temporal sequences and more garment instances to further illustrate the behavior of SCLTrack under diverse non-rigid configurations.