JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

GarmentTracking: Category-Level Garment Pose Tracking

Han Xue[1,2], Wenqiang Xu[2] , Jieyi Zhang[2], Tutian Tang[2], Yutong Li[2],
Wenxin Du[2],Ruolin Ye[3], Cewu Lu[†1,2]

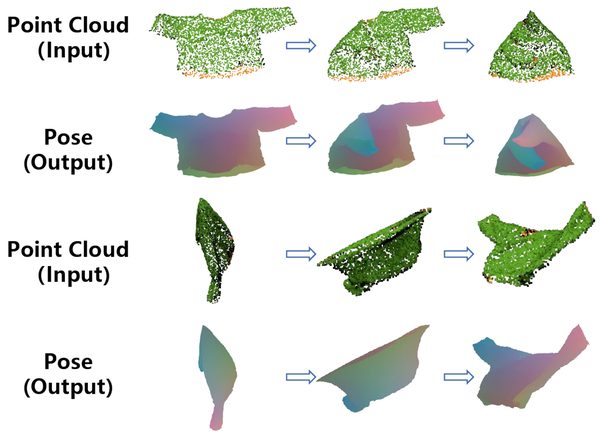
CVPR 2023 THU-PM-060

[1]Shanghai Qi Zhi institute [2]Shanghai Jiao Tong University [3]Cornell University

GarmentTracking: Category-Level Garment Pose Tracking

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Task Definition



Category level Garment Pose Tracking. We focus on the pose tracking problem in garment manipulation (e.g. flattening, folding). In this setting, we do not have the priors of the human body like previous works for clothed humans.

Challenges for Garment Perception

- Infinite DoF
- Severe Self-Occlusion
- Thin Structure

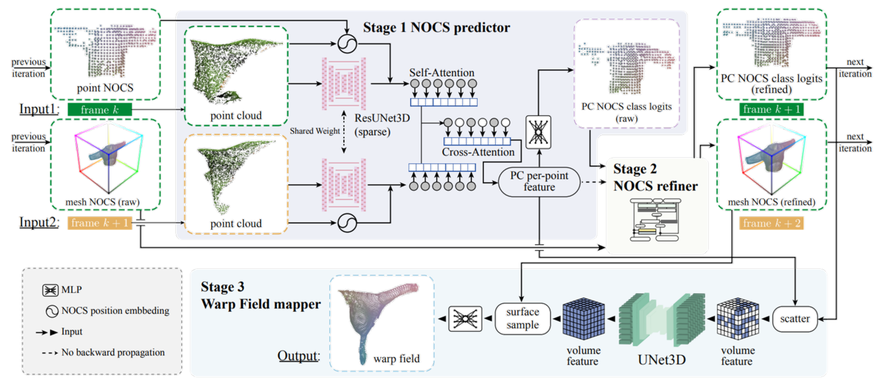
Challenges for Tracking Problem

- How to fuse inter-frame geometry and correspondence information?
- How to make the tracking prediction robust to pose estimation errors?
- How to achieve tracking in real-time?

Dataset: VR-Folding

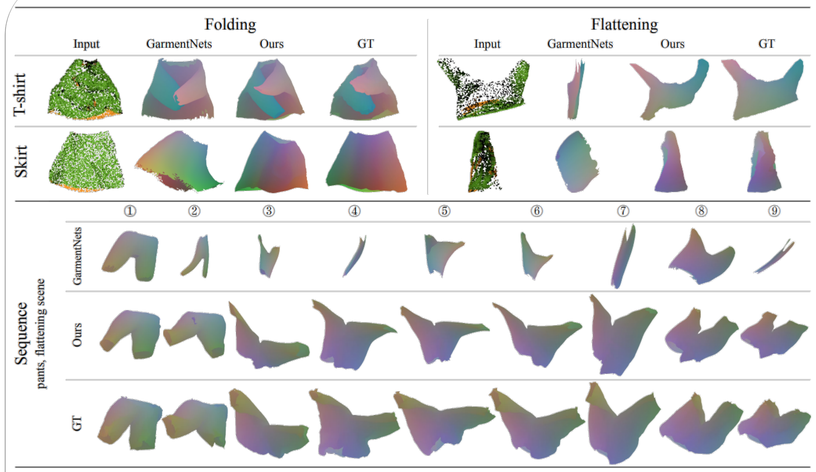
VR-Folding. We first create a real-time VR-based recording system named *VR-Garment*. Then the volunteer can manipulate the garment in a simulator through the VR interface. With *VR-Garment*, we build a large-scale garment manipulation dataset called *VR-Folding*. Our tasks include flattening and folding, which contain much complex garment configurations.

Garment Tracking Pipeline



The overview of GarmentTracking. Given the per-point NOCS coordinate of the first frame and a rough canonical shape (mesh NOCS), our tracking method takes two frames of the partial point cloud as input. In stage 1, the NOCS predictor will generate an inter-frame fusion feature and predict raw NOCS coordinates. In stage 2, the NOCS refiner will refine the NOCS coordinates and the canonical shape simultaneously. In stage 3, the warp field mapper will predict the warp field which maps from canonical space to task space.

Results

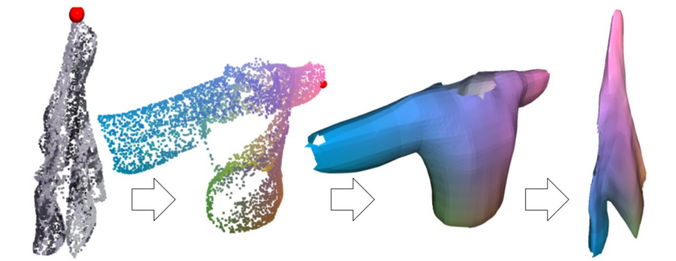


The qualitative results of pose estimation for unseen instances in VR-Folding dataset. In the long sequence tracking (shown in the lower part), our prediction still keeps high consistency with GT, while GarmentNets outputs a series of meshes that lack stability.

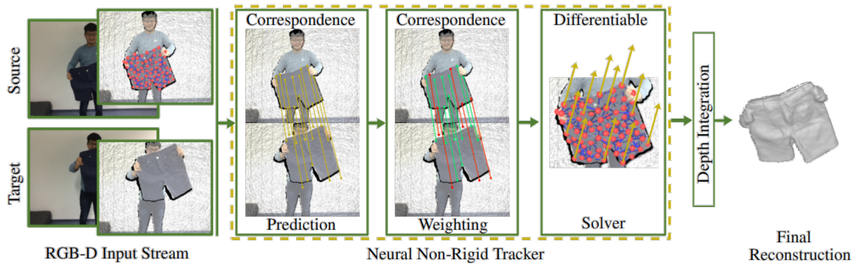
Type	Method	Init.	Folding $A_{3cm}\uparrow$	Folding $A_{5cm}\uparrow$	Folding $D_{corr}\downarrow$	Folding $D_{chamf}\downarrow$	Folding $D_{nocs}\downarrow$	Flattening $A_{5cm}\uparrow$	Flattening $A_{10cm}\uparrow$	Flattening $D_{corr}\downarrow$	Flattening $D_{chamf}\downarrow$	Flattening $D_{nocs}\downarrow$
Shirt	GarmentNets	N/A	0.8%	21.5%	6.40	1.58	0.221	13.2%	59.4%	10.54	3.54	0.135
	Ours	GT	**29.8%**	85.8%	**3.88**	**1.16**	**0.051**	**30.7%**	**83.4%**	**8.63**	**1.78**	**0.105**
	Ours	Pert.	29.0%	**85.9%**	3.88	1.18	0.052	25.4%	81.6%	8.94	1.85	0.109
Pants	GarmentNets	N/A	16.2%	69.5%	4.43	1.30	0.162	1.5%	42.4%	12.54	4.19	0.185
	Ours	GT	**47.3%**	**94.0%**	**3.26**	**1.07**	**0.039**	**31.3%**	**78.2%**	**8.97**	**1.64**	**0.113**
	Ours	Pert.	42.8%	93.6%	3.35	1.10	0.039	30.7%	76.9%	9.55	2.71	0.143
Top	GarmentNets	N/A	10.3%	53.8%	5.19	1.51	0.148	21.6%	57.6%	9.98	2.13	0.174
	Ours	GT	**37.9%**	**85.9%**	**3.75**	**0.99**	**0.051**	**36.5%**	**69.0%**	**9.41**	**1.59**	**0.113**
	Ours	Pert.	36.6%	86.1%	3.76	1.00	0.051	33.5%	68.1%	9.61	1.62	0.116
Skirt	GarmentNets	N/A	1.1%	30.3%	6.95	1.89	0.239	0.1%	7.9%	18.48	5.99	0.287
	Ours	GT	**23.5%**	**71.3%**	**4.61**	**1.33**	**0.060**	**5.4%**	**39.4%**	**16.09**	**2.02**	**0.199**
	Ours	Pert.	22.8%	70.6%	4.72	1.36	0.060	2.3%	35.5%	16.55	2.15	0.207

The quantitative results in VR-Folding dataset. In general, our method outperforms GarmentNets in all metrics by a large margin. On the challenging A_{3cm} metric in Folding task and A_{5cm} in Flattening task, GarmentNets has very low performance (e.g. 0.8% in Shirt Folding), while our method achieves much higher scores (e.g. 29.0% in Shirt Folding), which proves that our method can generate more accurate predictions in videos compared to GarmentNets. Our method also outperforms GarmentNets on mean correspondence distance and chamfer distance, which proves that our method can do well in both pose estimation and surface reconstruction tasks. Even with perturbation on first-frame poses (Ours with Pert. in Tab. 1), our method only shows minor performance loss (e.g. 37.9% → 36.6% in Top Folding) compared to using ground-truth as first-frame pose.

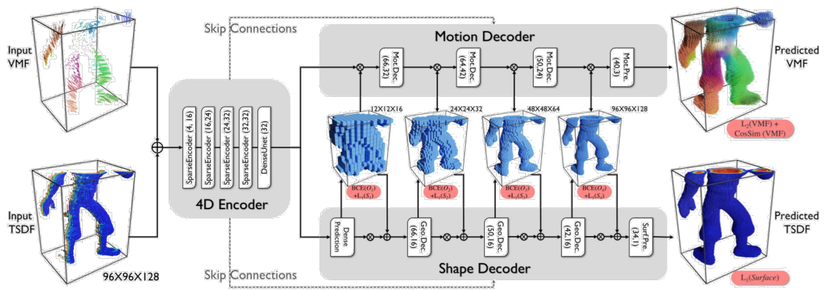
Task Definition



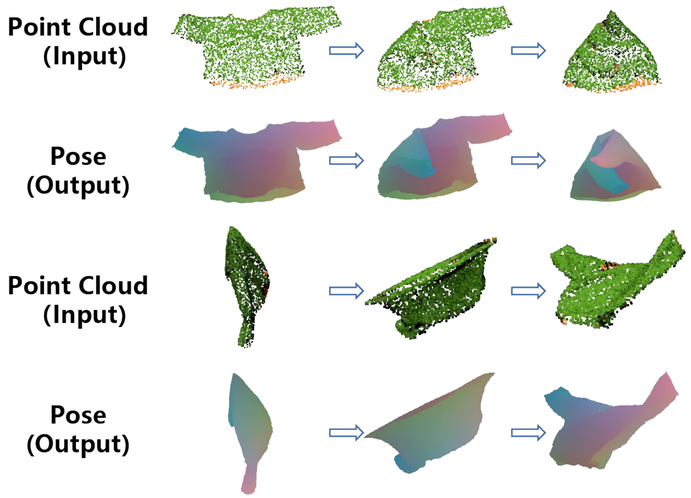
Garment Pose Estimation[1]



Non-Rigid Tracking Task[2]



Non-Rigid 4D Reconstruction[3]



Category level Garment Pose Tracking (Ours)

We focus on the pose tracking problem in garment manipulation (e.g. flattening, folding). In this setting, we do not have the priors of the human body like previous works for clothed humans.

Challenges for Garment Perception

- Infinite DoF
- Severe Self-Occlusion
- Thin Structure

Challenges for Tracking Problem

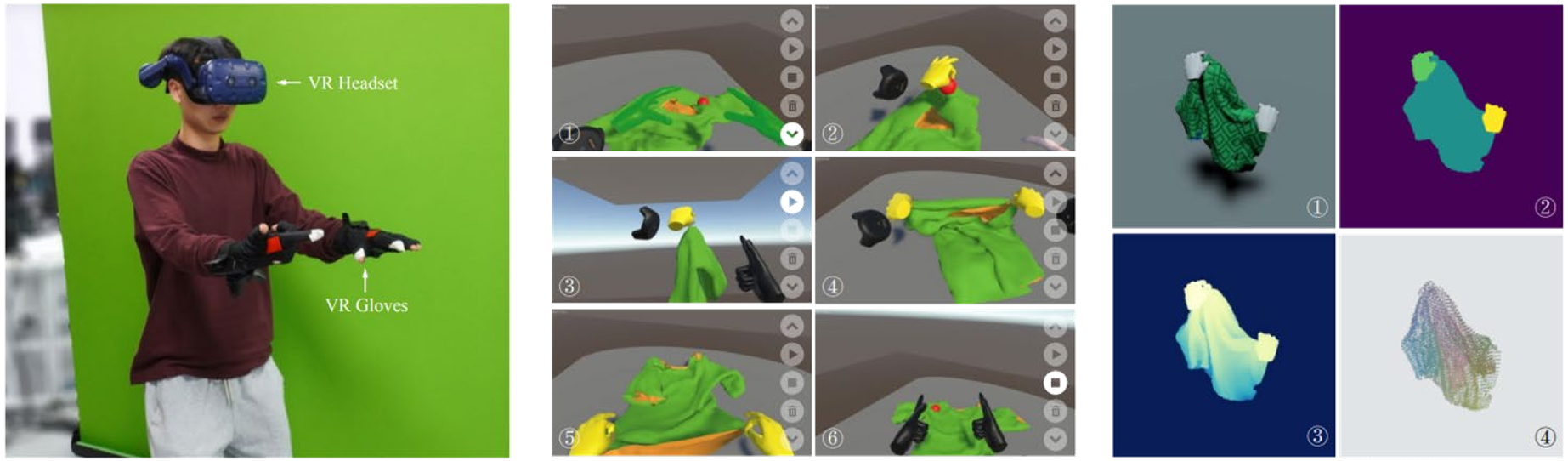
- How to fuse inter-frame geometry and correspondence information?
- How to make the tracking prediction robust to pose estimation errors?
- How to achieve tracking in real-time?

[1] Cheng Chi, et al. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3324–3333, 2021

[2] Aljaz Bozic, et al. Neural non-rigid tracking. Advances in Neural Information Processing Systems, 33:18727–18737, 2020.

[3] Yang Li, et al. 4dcomplete: Non-rigid motion estimation beyond the observable surface. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 12706–12716, 2021.

VR-Garment Platform



(a) The recording environment (b) The steps to record flattening (first-person view) (c) The re-rendered data

The pipeline of our Virtual Realty recording system (VR-Garment). (a) A volunteer needs to put on a VR headset and VR gloves. (b) By following the guidance of a specially designed UI, the volunteer begins to collect data efficiently. (c) After recording, we re-render multi-view RGB-D images with Unity [6] and obtain masks and deformed garment meshes with NOCS labels.

VR-Folding Dataset

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

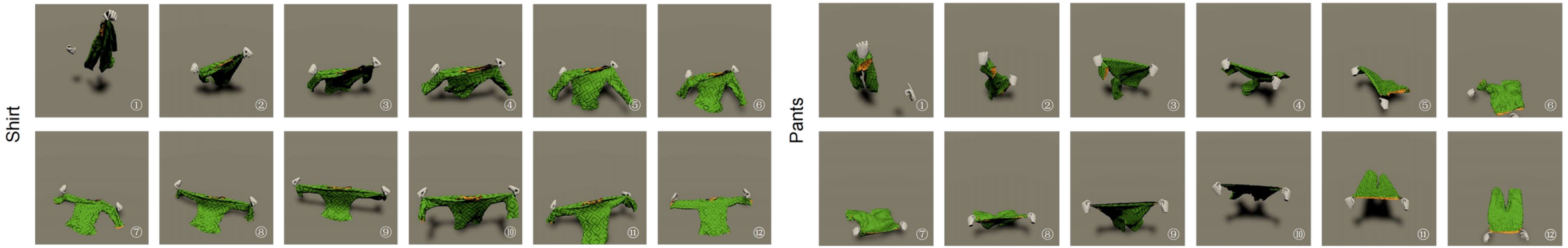


Figure 1. The examples of **Flattening** task.

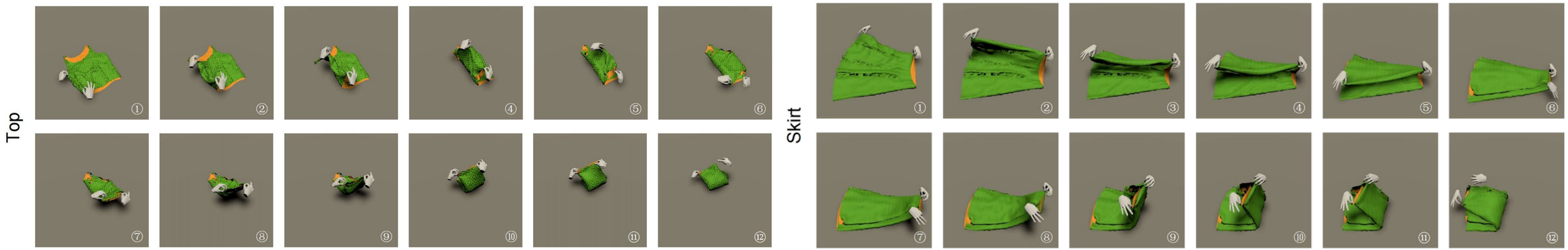
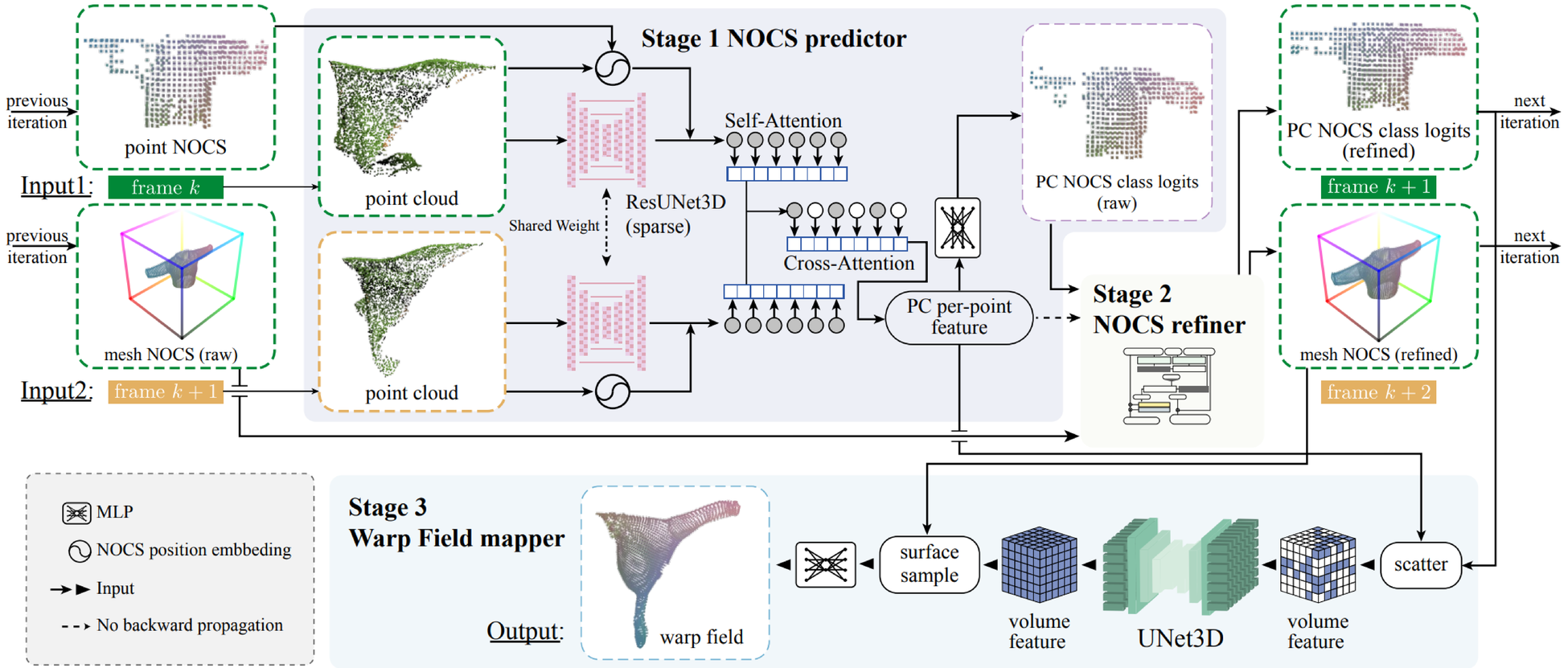


Figure 2. The examples of **Folding** task.

Garment Tracking

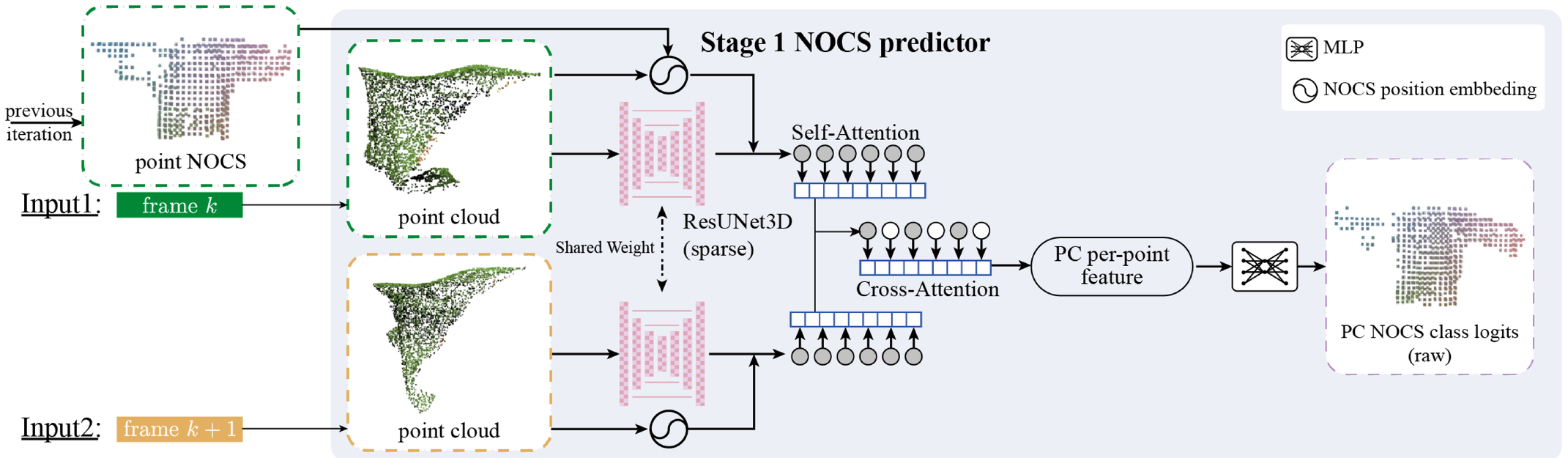
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA



The overview of GarmentTracking. Given the per-point NOCS coordinate of the first frame and a rough canonical shape (mesh NOCS), our tracking method takes two frames of the partial point cloud as input. In stage 1, the NOCS predictor will generate an inter-frame fusion feature and predict raw NOCS coordinates. In stage 2, the NOCS refiner will refine the NOCS coordinates and the canonical shape simultaneously. In stage 3, the warp field mapper will predict the warp field which maps from canonical space to task space.

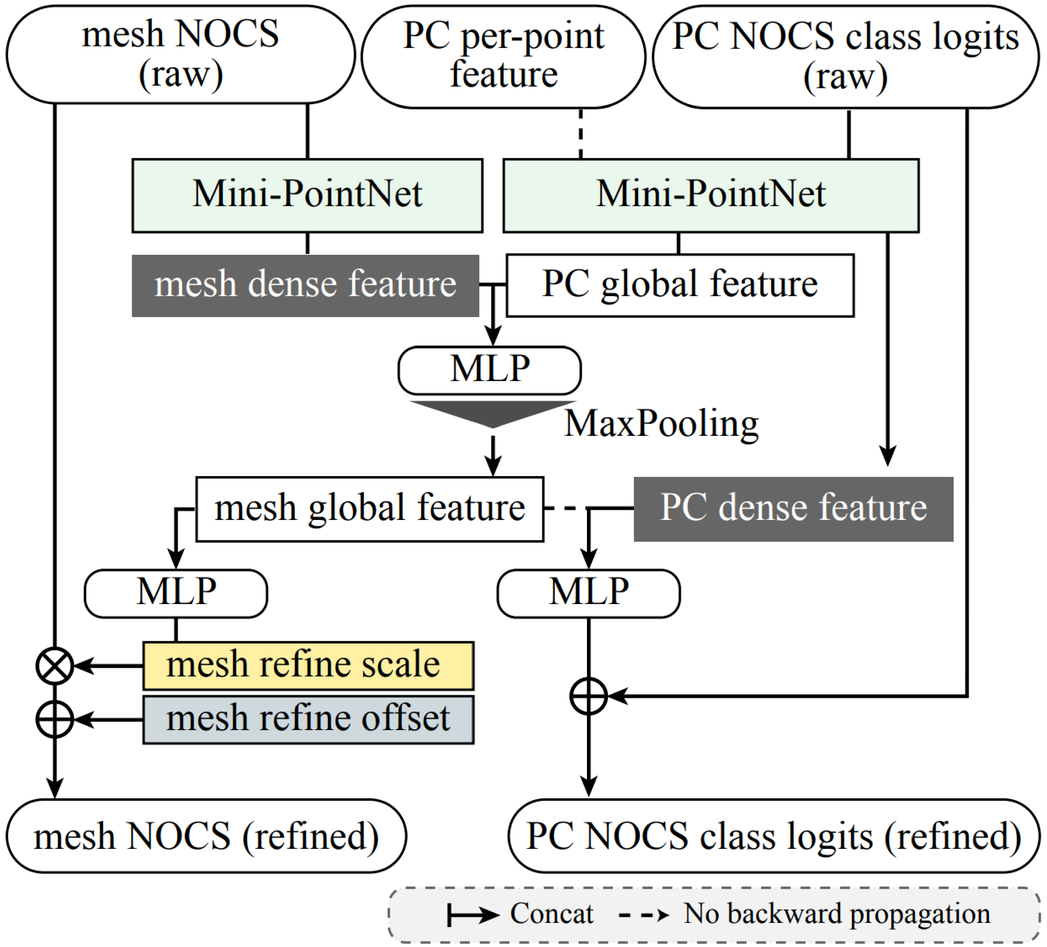
Garment Tracking Stage 1

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA



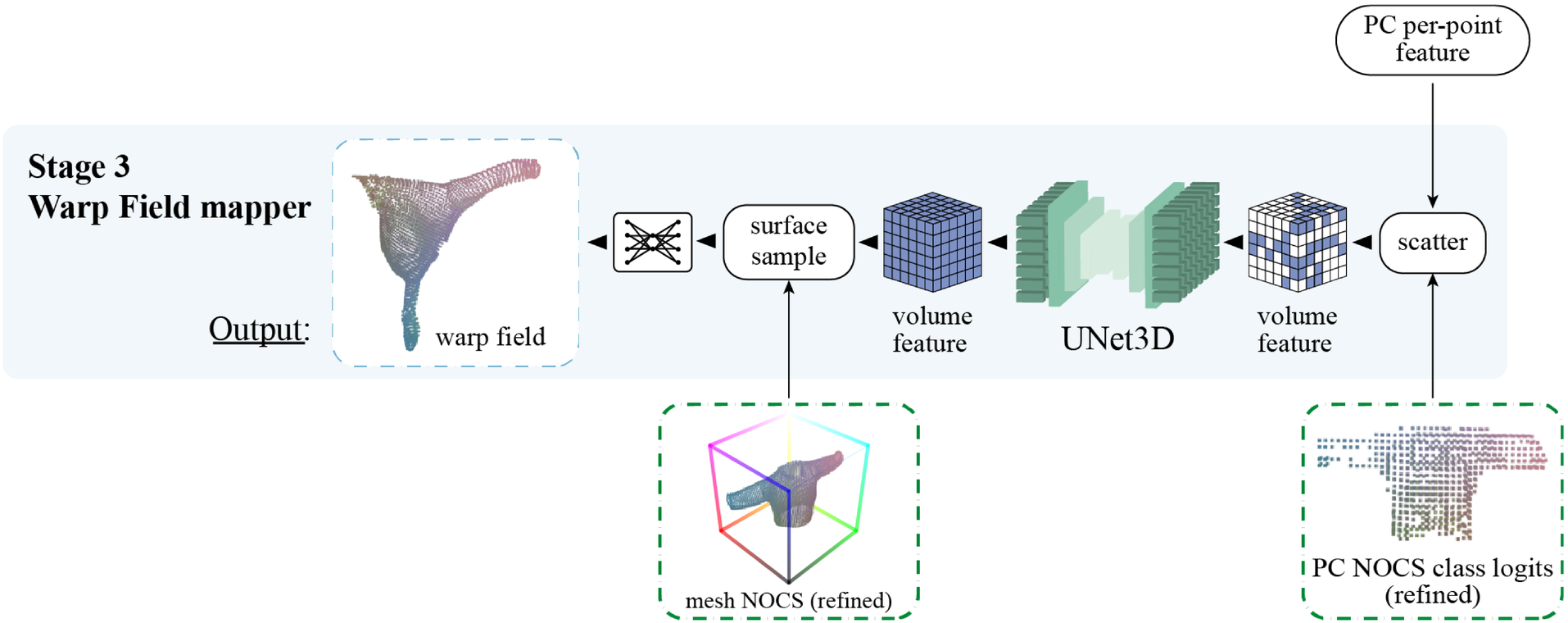
Garment Tracking Stage 2

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA



Stage 2 NOCS Refiner

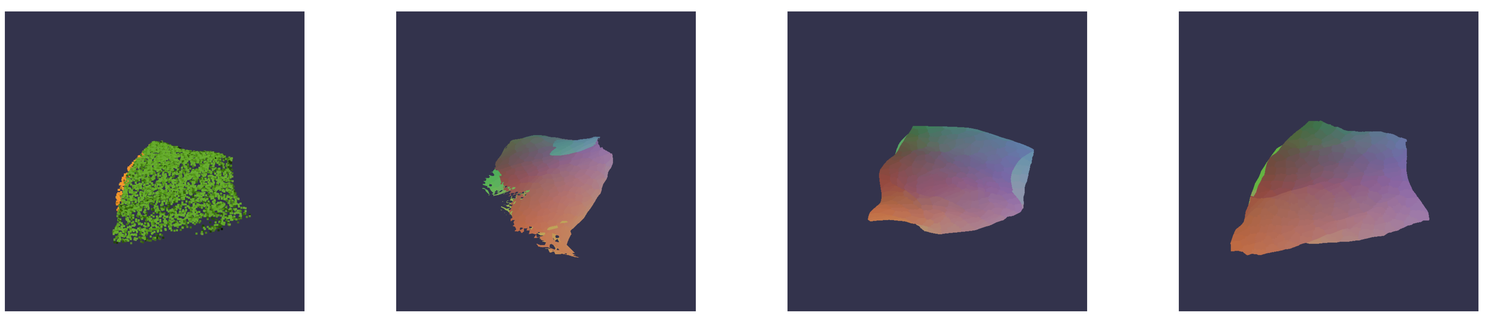
Garment Tracking Stage 3



Experiment Result

Type	Method	Init.	Folding					Flattening				
			$A_{3cm} \uparrow$	$A_{5cm} \uparrow$	$D_{corr} \downarrow$	$D_{chamf} \downarrow$	$D_{nocs} \downarrow$	$A_{5cm} \uparrow$	$A_{10cm} \uparrow$	$D_{corr} \downarrow$	$D_{chamf} \downarrow$	$D_{nocs} \downarrow$
Shirt	GarmentNets	N/A	0.8%	21.5%	6.40	1.58	0.221	13.2%	59.4%	10.54	3.54	0.135
	Ours	GT	**29.8%**	85.8%	**3.88**	**1.16**	**0.051**	**30.7%**	**83.4%**	**8.63**	**1.78**	**0.105**
	Ours	Pert.	29.0%	**85.9%**	3.88	1.18	0.052	25.4%	81.6%	8.94	1.85	0.109
Pants	GarmentNets	N/A	16.2%	69.5%	4.43	1.30	0.162	1.5%	42.4%	12.54	4.19	0.185
	Ours	GT	**47.3%**	**94.0%**	**3.26**	**1.07**	**0.039**	**31.3%**	**78.2%**	**8.97**	**1.64**	**0.113**
	Ours	Pert.	42.8%	93.6%	3.35	1.10	0.039	30.7%	76.9%	9.55	2.71	0.143
Top	GarmentNets	N/A	10.3%	53.8%	5.19	1.51	0.148	21.6%	57.6%	9.98	2.13	0.174
	Ours	GT	**37.9%**	**85.9%**	**3.75**	**0.99**	**0.051**	**36.5%**	**69.0%**	**9.41**	**1.59**	**0.113**
	Ours	Pert.	36.6%	86.1%	3.76	1.00	0.051	33.5%	68.1%	9.61	1.62	0.116
Skirt	GarmentNets	N/A	1.1%	30.3%	6.95	1.89	0.239	0.1%	7.9%	18.48	5.99	0.287
	Ours	GT	**23.5%**	**71.3%**	**4.61**	**1.33**	**0.060**	**5.4%**	**39.4%**	**16.09**	**2.02**	**0.199**
	Ours	Pert.	22.8%	70.6%	4.72	1.36	0.060	2.3%	35.5%	16.55	2.15	0.207

Quantitative results on VR-Folding dataset



Input GarmentNets Ours GT

Qualitative results on VR-Folding dataset

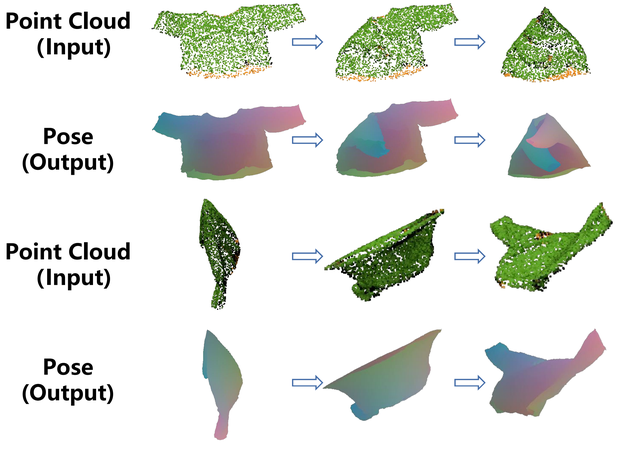
GarmentTracking: Category-Level Garment Pose Tracking

Han Xue[1,2], Wenqiang Xu[2] , Jieyi Zhang[2], Tutian Tang[2], Yutong Li[2], Wenxin Du[2],Ruolin Ye[3], Cewu Lu[†1,2]
[1]Shanghai Qi Zhi institute [2]Shanghai Jiao Tong University [3]Cornell University

garment-tracking.robotflow.ai

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Task Definition



Category level Garment Pose Tracking. We focus on the pose tracking problem in garment manipulation (e.g. flattening, folding). In this setting, we do not have the priors of the human body like previous works for clothed humans.

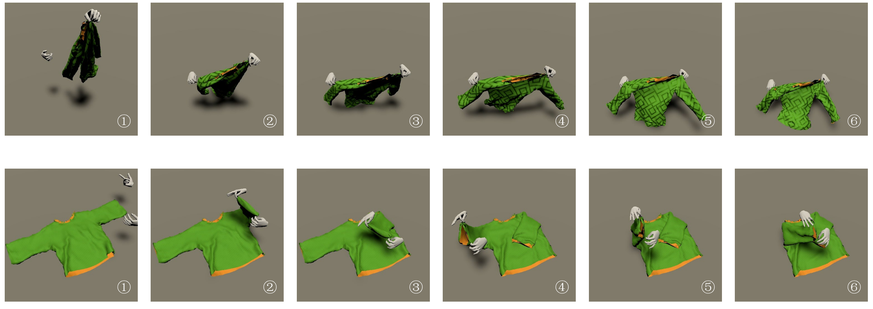
Challenges for Garment Perception

- Infinite DoF
- Severe Self-Occlusion
- Thin Structure

Challenges for Tracking Problem

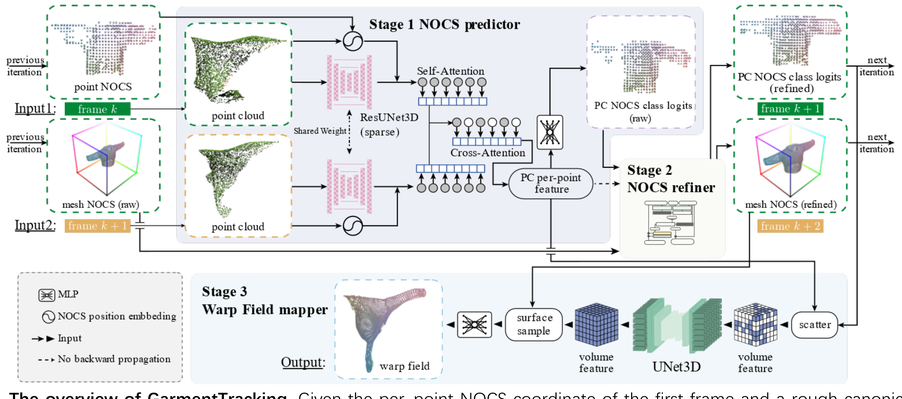
- How to fuse inter-frame geometry and correspondence information?
- How to make the tracking prediction robust to pose estimation errors?
- How to achieve tracking in real-time?

Dataset: VR-Folding



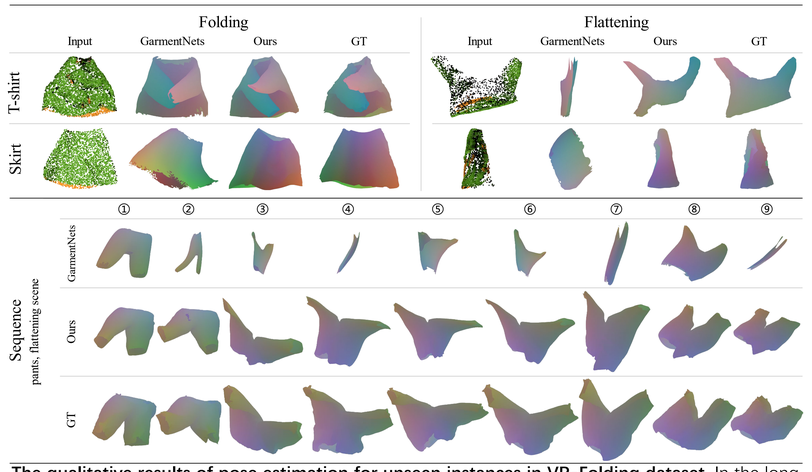
VR-Folding. We first create a real-time VR-based recording system named *VR-Garment*. Then the volunteer can manipulate the garment in a simulator through the VR interface. With *VR-Garment*, we build a large-scale garment manipulation dataset called *VR-Folding*. Our tasks include flattening and folding, which contain much complex garment configurations.

Garment Tracking Pipeline



The overview of GarmentTracking. Given the per-point NOCS coordinate of the first frame and a rough canonical shape (mesh NOCS), our tracking method takes two frames of the partial point cloud as input. In stage 1, the NOCS predictor will generate an inter-frame fusion feature and predict raw NOCS coordinates. In stage 2, the NOCS refiner will refine the NOCS coordinates and the canonical shape simultaneously. In stage 3, the warp field mapper will predict the warp field which maps from canonical space to task space.

Results



The qualitative results of pose estimation for unseen instances in VR-Folding dataset. In the long sequence tracking (shown in the lower part), our prediction still keeps high consistency with GT, while GarmentNets outputs a series of meshes that lack stability.

Type	Method	Init.	Folding					Flattening				
			$A_{3cm}\uparrow$	$A_{5cm}\uparrow$	$D_{corr}\downarrow$	$D_{chamf}\downarrow$	$D_{nocs}\downarrow$	$A_{5cm}\uparrow$	$A_{10cm}\uparrow$	$D_{corr}\downarrow$	$D_{chamf}\downarrow$	$D_{nocs}\downarrow$
Shirt	GarmentNets	N/A	0.8%	21.5%	6.40	1.58	0.221	13.2%	59.4%	10.54	3.54	0.135
	Ours	GT	**29.8%**	85.8%	**3.88**	**1.16**	**0.051**	**30.7%**	**83.4%**	**8.63**	**1.78**	**0.105**
	Ours	Pert.	29.0%	**85.9%**	3.88	1.18	0.052	25.4%	81.6%	8.94	1.85	0.109
Pants	GarmentNets	N/A	16.2%	69.5%	4.43	1.30	0.162	1.5%	42.4%	12.54	4.19	0.185
	Ours	GT	**47.3%**	**94.0%**	**3.26**	**1.07**	**0.039**	**31.3%**	**78.2%**	**8.97**	**1.64**	**0.113**
	Ours	Pert.	42.8%	93.6%	3.35	1.10	0.039	30.7%	76.9%	9.55	2.71	0.143
Top	GarmentNets	N/A	10.3%	53.8%	5.19	1.51	0.148	21.6%	57.6%	9.98	2.13	0.174
	Ours	GT	**37.9%**	**85.9%**	**3.75**	**0.99**	**0.051**	**36.5%**	**69.0%**	**9.41**	**1.59**	**0.113**
	Ours	Pert.	36.6%	86.1%	3.76	1.00	0.051	33.5%	68.1%	9.61	1.62	0.116
Skirt	GarmentNets	N/A	1.1%	30.3%	6.95	1.89	0.239	0.1%	7.9%	18.48	5.99	0.287
	Ours	GT	**23.5%**	**71.3%**	**4.61**	**1.33**	**0.060**	**5.4%**	**39.4%**	**16.09**	**2.02**	**0.199**
	Ours	Pert.	22.8%	70.6%	4.72	1.36	0.060	2.3%	35.5%	16.55	2.15	0.207

The quantitative results in VR-Folding dataset. In general, our method outperforms GarmentNets in all metrics by a large margin. On the challenging A_{3cm} metric in Folding task and A_{5cm} in Flattening task, GarmentNets has very low performance (e.g. 0.8% in Shirt Folding), while our method achieves much higher scores (e.g. 29.0% in Shirt Folding), which proves that our method can generate more accurate predictions in videos compared to GarmentNets. Our method also outperforms GarmentNets on mean correspondence distance and chamfer distance, which proves that our method can do well in both pose estimation and surface reconstruction tasks. Even with perturbation on first-frame poses (Ours with Pert. in Tab. 1), our method only shows minor performance loss (e.g. 37.9% → 36.6% in Top Folding) compared to using ground-truth as first-frame pose.