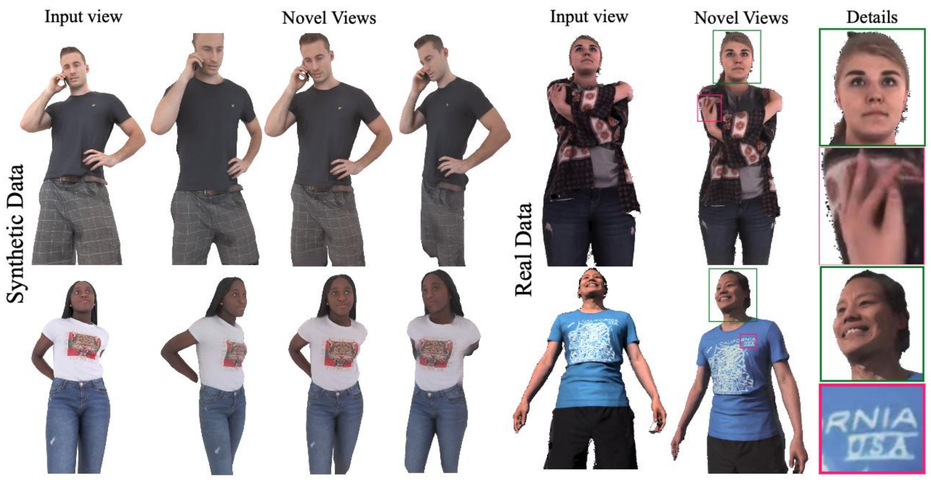
Free-Viewpoint RGB-D Human Performance Capture and Rendering

Phong Nguyen[1,2], Nikolaos Sarafianos[2], Christoph Lassner[2], Janne Heikkilä[1], Tony Tung[2]

[1] University of Oulu, Finland, [2] Meta Reality Labs Research, Sausalito

ECCV TEL AVIV 2022

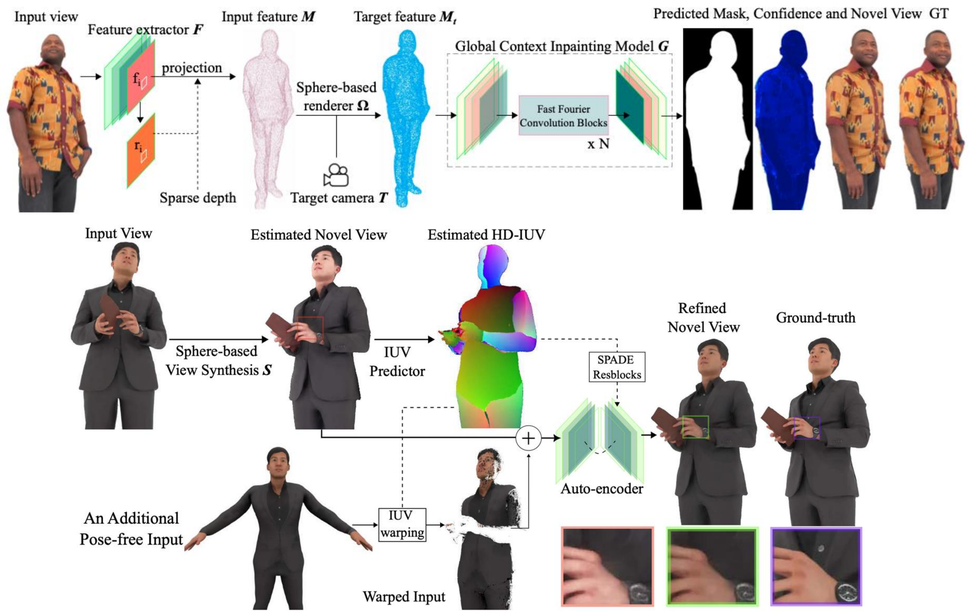
Generalizable Human View Synthesis



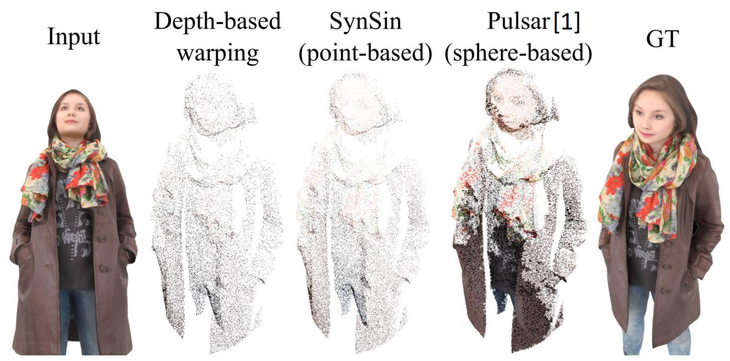
Contributions:

- Robust sphere-based synthesis network that generalizes to multiple identities without per-human optimization
- Refinement module that enhances the self-occluded regions of the initial estimated novel views
- State-of-the-art results on dynamic humans wearing various clothes, or accessories with varying of facial expressions of synthetic and real-captured data

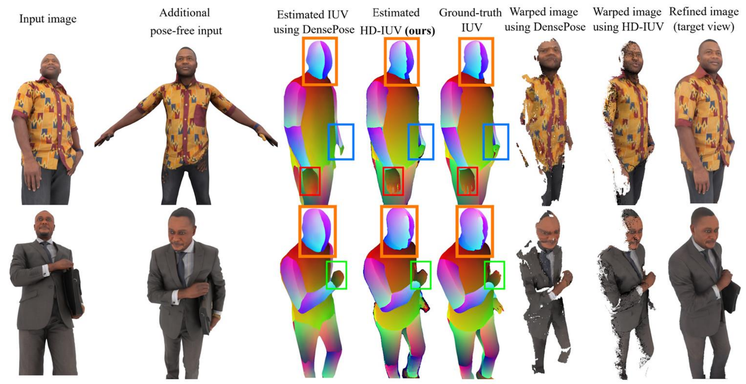
Human View Synthesis Network (HVS-Net)



Point vs Sphere-based Approaches



Dense Correspondence Visualization



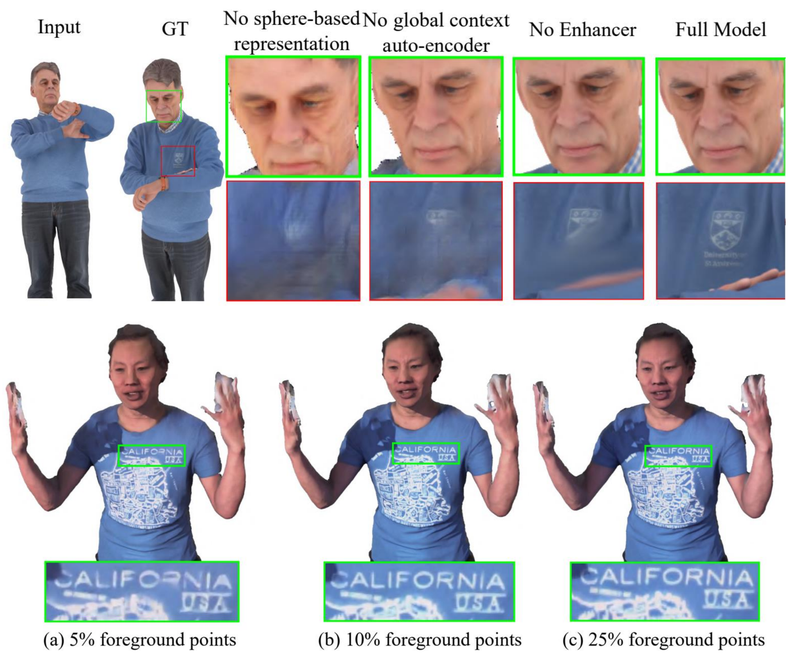
Our proposed HD-IUV representation covers the human body including clothing, captures facial and hand details with high accuracy, and results in less distorted renderings in the target view

Quantitative Results

Method	RenderPeople (static)			RenderPeople (animated)			Real 3dMD Data		
	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑	LPIPS↓	SSIM↑	PSNR↑
LookingGood†[2]	0.24	0.925	25.32	0.25	0.912	24.53	0.29	0.863	25.12
SynSin†[3]	0.31	0.851	24.18	0.35	0.937	23.64	0.35	0.937	22.18
SynSin[3]	0.52	0.824	22.45	0.55	0.853	20.86	0.65	0.819	19.92
HVS-Net (w/o Enhancer)	0.18	0.986	28.54	0.19	0.926	26.24	0.20	0.910	26.25
HVS-Net†	**0.14**	**0.986**	**28.56**	**0.17**	**0.958**	27.41	**0.20**	**0.918**	**26.47**
HVS-Net	0.15	**0.986**	28.54	**0.17**	0.955	**27.45**	**0.20**	**0.918**	**26.47**

Methods with a † symbol are using dense input depth. Both **HVS-Net** and **HVS-Net†** outperforms other baselines both qualitative and quantitatively.

Ablation Studies: Architecture Design & Varying Depth Sparsity



Method Variant	LPIPS↓	SSIM↑	PSNR↑
No Sphere Repres.	0.22	0.934	26.15
No Global Context	0.21	0.954	26.82
No Enhancer	0.18	0.967	27.92
HVS-Net (full)	**0.15**	**0.986**	**28.54**

Input depth (%)	Run-time↑ (fps)	LPIPS↓	SSIM↑	PSNR↑
5	**25**	0.17	0.985	28.27
10	22	0.15	0.986	28.54
25	21	**0.14**	0.986	28.55
100	20	0.14	0.986	**28.56**

Real-time Application:

- User passes a target viewpoint and we generate novel views across the trajectory path from the current view to the target user view.
- Runs at 30fps at 512x512 resolution on an NVIDIA RTX 3080.

References

[1]Pulsar: Efficient sphere-based rendering, CVPR 2021.
[2]LookinGood: Enhancing performance capture with real-time neural re-rendering, ACM ToG 2018.
[3]Synsin: End-to-end view synthesis from a single image, CVPR 2020.