3DV 2021

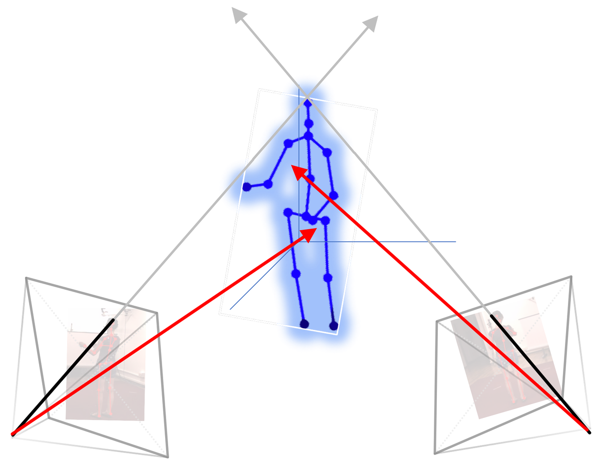
High Fidelity 3D Reconstructions with Limited Physical Views

[1] Carnegie Mellon University
[2] Apple Inc.
[3] The University of Adelaide

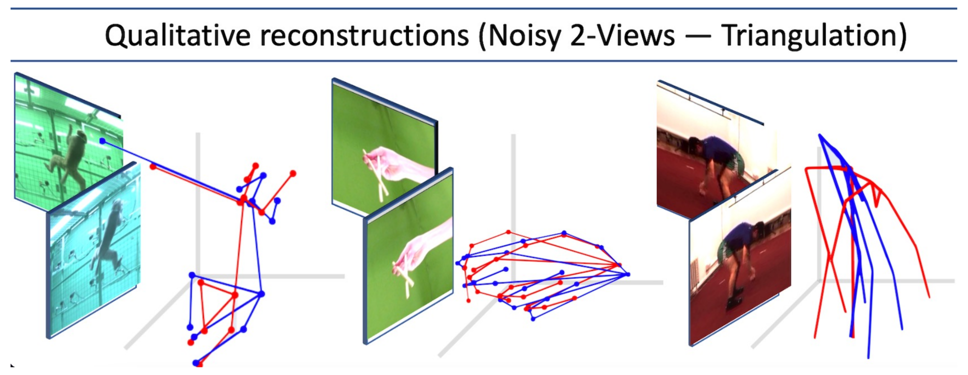
Mosam Dabhi[1,2] Chaoyang Wang[1] Kunal Saluja[2] Laszlo Jeni[1] Ian Fasel[2] Simon Lucey[1,3]

Overview

Two views are not enough for triangulation



Triangulation from two views



Resulting two-view reconstructions on monkeys, hands, bodies

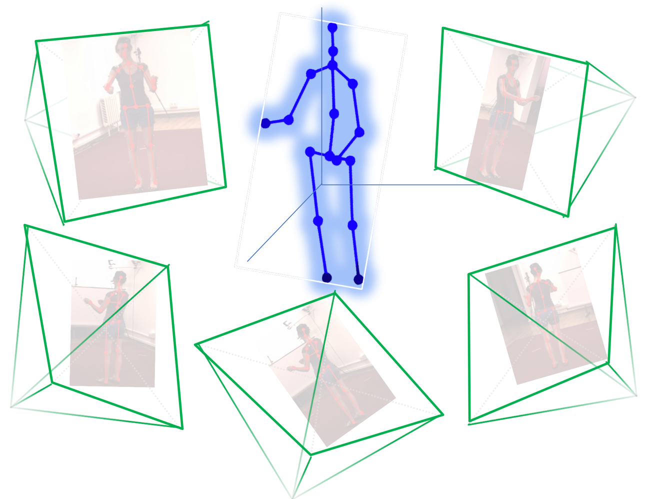
In principle, two views should be enough to triangulate a point. However, any imperfections in 2D keypoints or calibration leads to poor reconstructions.

There are no constraints for reconstructing the points and they could end up arbitrarily anywhere.

Large multi-view rigs (shown below) enables the usage of accurate camera calibration and multiple views to minimize error on each point. However, that could lead to immense cost and complexity.

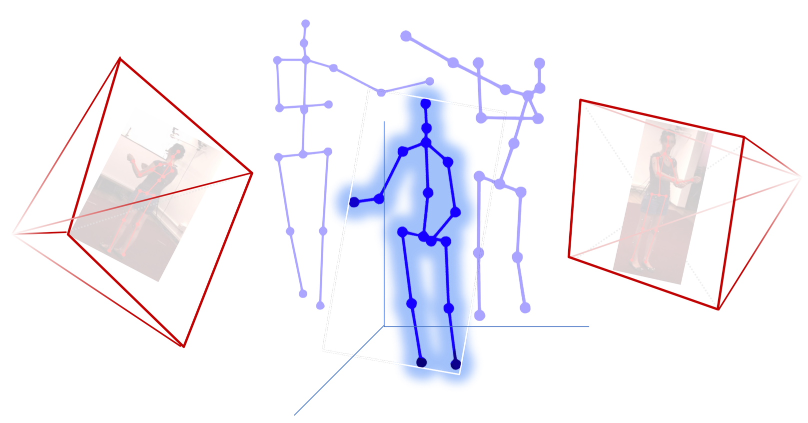
Approach: Two views can be enough!

Multiple views + accurate camera calibration



Large multi-view rig uses multiple observations (with outlier rejection) to minimize error for each point.
(Still, it does not enforce any constraints on the overall shape)

Two uncalibrated views + neural shape prior



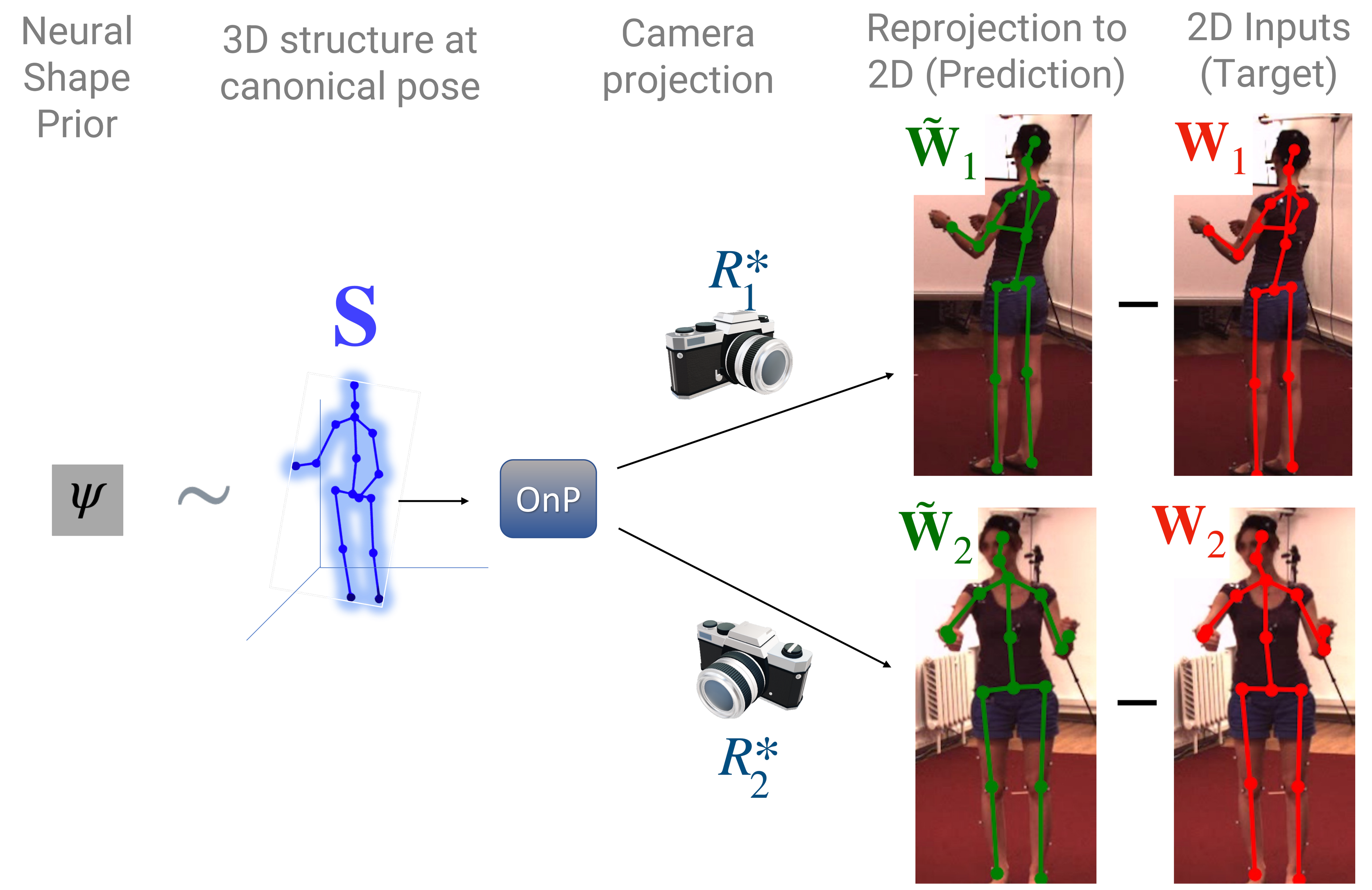
Our approach: Instead of more cameras, we add a neural prior to constrain the shape (the set of 3D points) to lie on a manifold.

This allows us to combine multiple observations even though the object is deforming, while only leveraging only two physical views at any observation.

https://sites.google.com/view/high-fidelity-3d-neural-prior

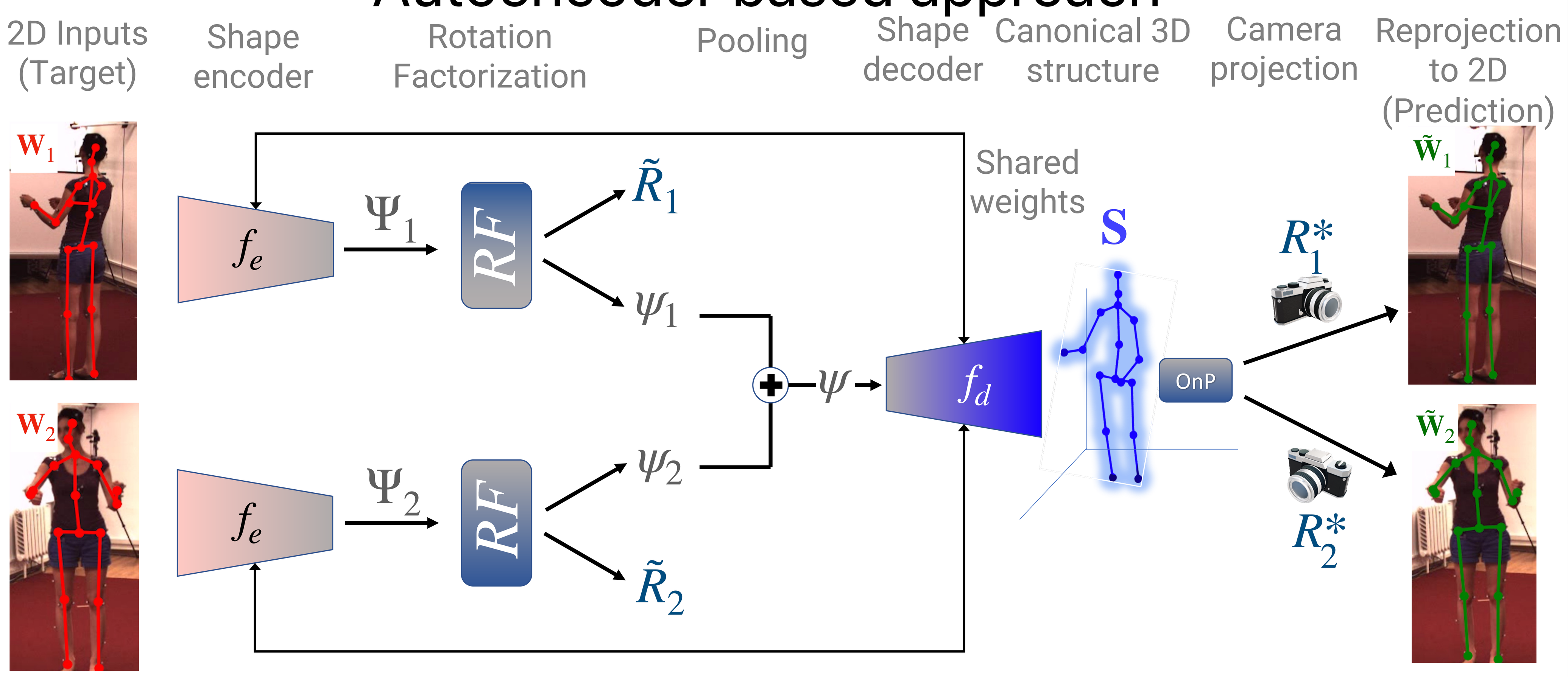
Method

Statistical Shape Prior



- The 3D structure, $\mathbf{S}$ is drawn from a statistical shape distribution using neural shape priors and projected to 2 views using the Orthographic-N-Point (OnP).
- Parameters of the shape distribution are adapted by minimizing the predicted and groundtruth (input) 2D projections.
- $\mathbf{S}$, R^*, and $\mathbf{W}$ are recovered by constraining shapes from a shared neural model.

Autoencoder based approach

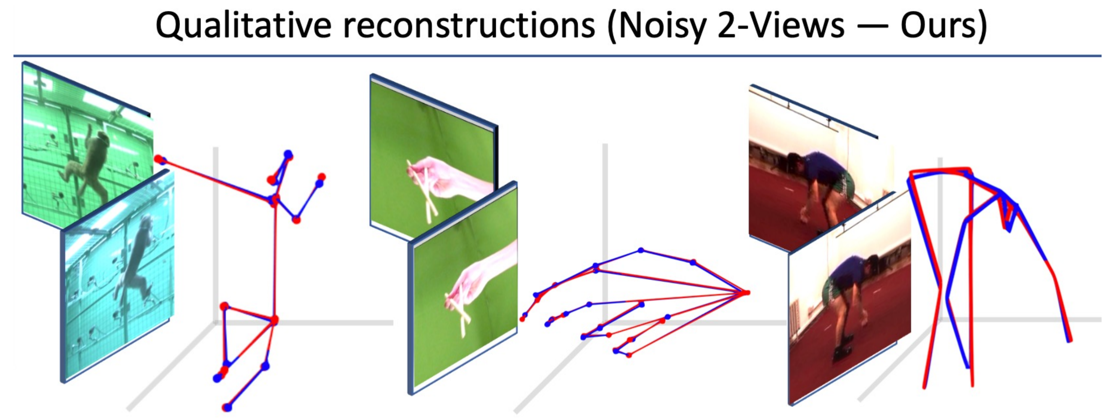


- Motivated by hierarchical sparse coding, network f_e extracts block sparse codes Ψ.
- The bottleneck (RF layer) extracts each block sparse code into camera matrix and unrotated vector sparse code.
- Codes are pooled and fed into the decoder f_d to generate canonicalized 3D structure $\mathbf{S}$.

Results

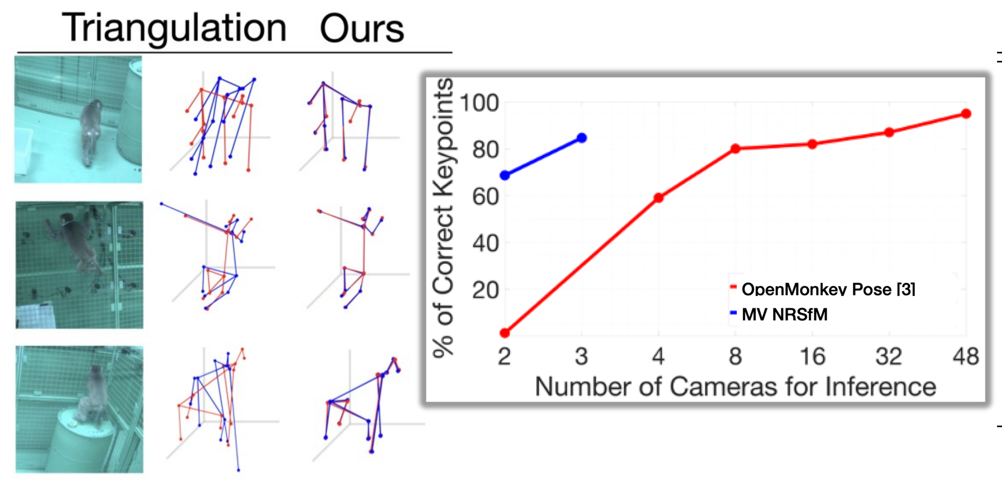
Robustness to calibration and noise on 2D keypoints

Multiple modalities



Resulting two-view reconstructions on monkeys, hands, bodies

Monkey dataset



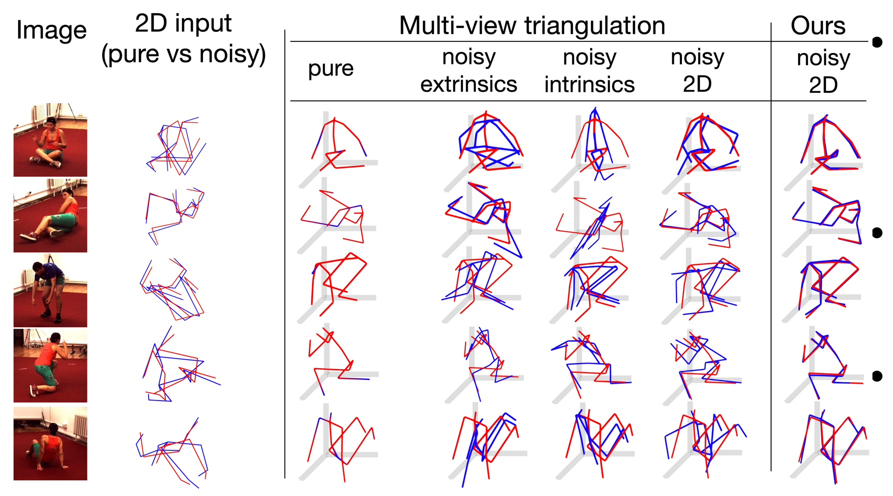
Method	PCK
2 Views	1.2%
4 Views	59%
8 Views	80%
16 Views	82%
32 Views	87%
48 Views	95%
2-views (ours)	**68.63%**
3-views (ours)	**84.63%**

- Our method with 3 views is comparable to 16+ views that utilizes iterative multi-view triangulation (TRNG).

Human dataset

	S1, S5, S6, S7, S8								
	Extrinsics Noise			Intrinsics Noise			2D keypoints Noise		
	$\sigma=0.1$	$\sigma=0.5$	$\sigma=0.9$	$\sigma=0.1$	$\sigma=0.5$	$\sigma=0.9$	$\sigma=15$	$\sigma=25$	$\sigma=35$
TRNG	65.49	131.66	145.94	69.57	188.63	234.47	70.08	114.06	154.41
2-Views (ours)	**30.53**						**54.22**	**65.74**	**77.82**

Robustness to camera calibration and 2D annotations noise for Human 3.6M dataset. Values are in **mm**.



- Our method is robust to both 2D keypoints and calibration noise.
- On top of 3D structure, we also return camera matrix.
- This should enable data collection with imperfect/no calibration.

Conclusion

- This work could open doors for wide-scale data collection setups, making the expensive and complex multi-view rigs obsolete.
- **Limitation**: Requires multiple non-rigid atemporal views to enforce the proposed neural shape prior during optimization.