MBW: Multiview Bootstrapping in the Wild

Mosam Dabhi Chaoyang Wang Tim Clifford Laszlo Jeni Ian Fasel Simon Lucey

NEURAL INFORMATION PROCESSING SYSTEMS

Few-shot labeling **in-the-wild** at scale

Prevalent categories

Long tail distribution categories (in-the-wild)

Unlike prevalent categories, annotated data for long tail distribution categories casually captured in the wild is largely out of reach!

Summary: MBW obtains high-fidelity **2D** and **3D landmark labels** for deformable object categories from videos with only **two or three uncalibrated**, **handheld** cameras moving **in the wild**. By leveraging neural priors, MBW carries out geometry based Out-of-Distribution (OOD) detection for data labeling at scale in a few-shot learning fashion.

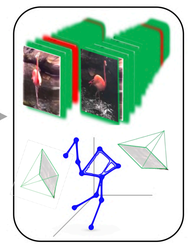
Problem setup

Capture: Casually using handheld cameras

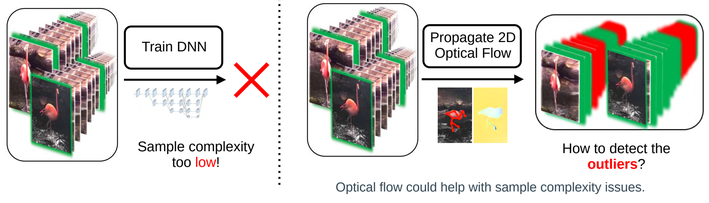
Input: 2 view videos (*calibration **not** required)*

Few-shot labels: About 10-15 frames

Output: 2D + 3D landmark predictions w/ cameras



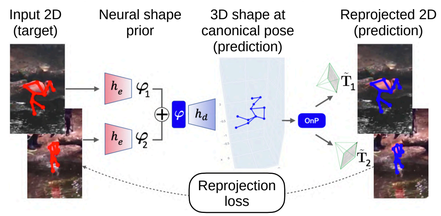
Sample complexity issues



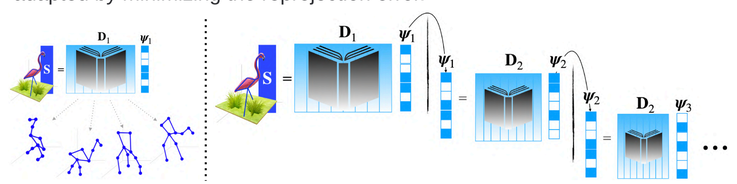
Optical flow could help with sample complexity issues.

Self-training by **neural shape priors**

We use neural shape priors with multi-view equivariance to detect the outliers.

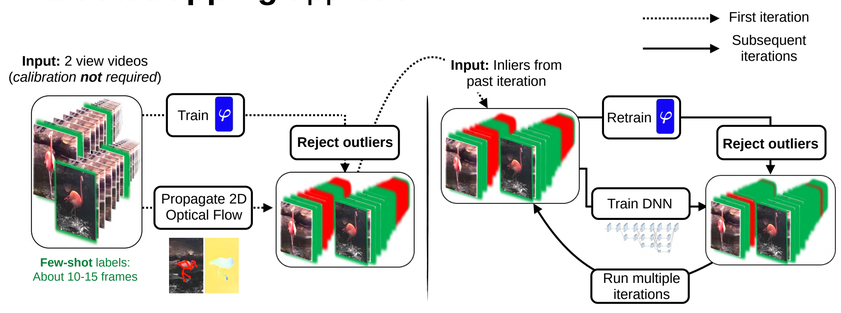


3D structure is drawn from a statistical shape distribution using neural shape priors and projected to 2 views using OnP. Parameters of this distribution are adapted by minimizing the reprojection error.



This distribution is learned by enforcing a shape prior on 3D structure, where it decomposes into an overcomplete hierarchical sparse dictionary.

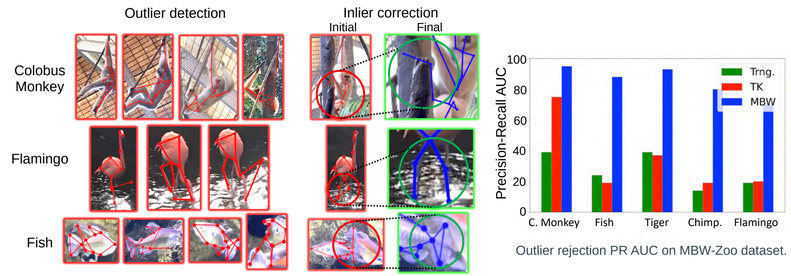
Bootstrapping approach



(**Dotted line**) The neural prior lifting network is initially trained with few-shot frames (shown as green images). A pre-trained optical flow network then propagates the initial labels through the video to generate additional 2D pseudo labels. Candidates that result in high reprojection error from the 3D lifting network are rejected as outliers (red).

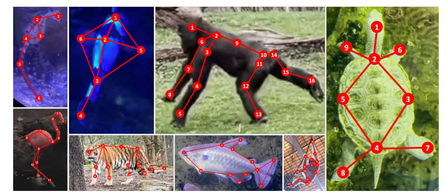
(**Solid line**) From here on, the label set is updated with inliers from the previous iteration, and is then used both to retrain the neural shape prior network and to train a 2D detector.

Geometry based **OOD Detection** in-the-wild



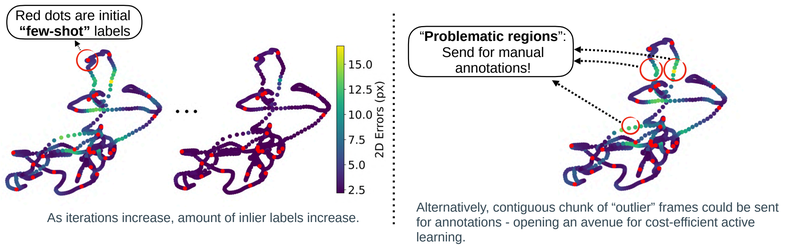
Outlier rejection PR AUC on MBW-Zoo dataset.

MBW-Zoo dataset



We release code + dataset consisting 2D/3D landmarks with bounding box labels.

Active learning via φ



As iterations increase, amount of inlier labels increase.

Alternatively, contiguous chunk of "outlier" frames could be sent for annotations - opening an avenue for cost-efficient active learning.

Benchmarking against groundtruth

Human3.6M	Nec ↑	Sho ↑	Elb ↑	Wri ↑	Hip ↑	Kne ↑	Ank ↑
Dong et al.	91.7	81.4	42.3	25.6	93.9	83.4	87.5
Jafarian et al.	89.6	48.3	29.7	20.5	29.8	34.9	60.7
Zheng & Park	93.2	**92.8**	67.3	49.6	**93.7**	**87.6**	89.5
MBW (Ours)	**96.8**	83.3	**78.1**	**69.8**	89.2	82.9	**92.9**

PCKh measure to test generalizability of 2D detection on unseen data of Human3.6M. We use just 2% of the labeled data compared to other approaches.

Multi-view methods	(mm) ↓
Martinez et al. (multi-view).	46.5
Pavlakos et al.	41.2
Kadkhodamohammadi & Padoy	39.4
Iskakov et al.	19.9
Reddy et al.	**17.5**
Ours (PA-MPJPE at 2%)	20.15

3D reconstruction accuracy of different methods on the Human3.6M dataset.

References

Dabhi et al. (2021). "High Fidelity 3D Reconstructions with Limited Physical Views." In 2021 International Conference on 3D Vision (3DV) (pp. 1301-1311). IEEE.

Teed, Z., & Deng, J. (2020). "Raft: Recurrent all-pairs field transforms for optical flow." In European conference on computer vision (pp. 402-419). Springer, Cham.

Sun et al. (2019). "Deep high-resolution representation learning for human pose estimation." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5693-5703).