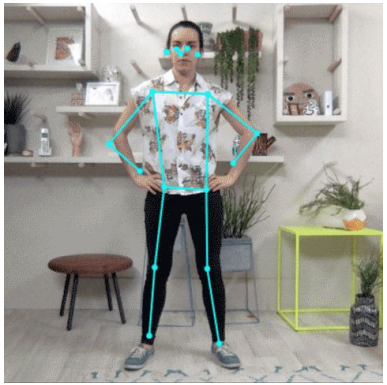
Light3DPose

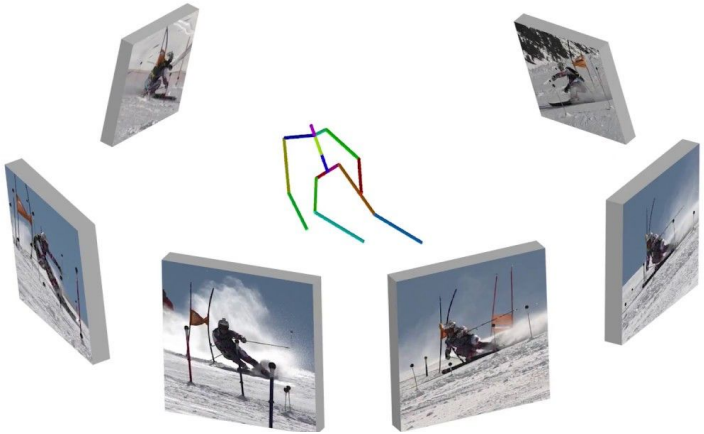
Real-time Multi-Person 3D Pose Estimation from Multiple Views

Alessio Elmi - Davide Mazzini - Pietro Tortella
{alessio, davide, pietro}@standard.ai

STANDARD

Introduction



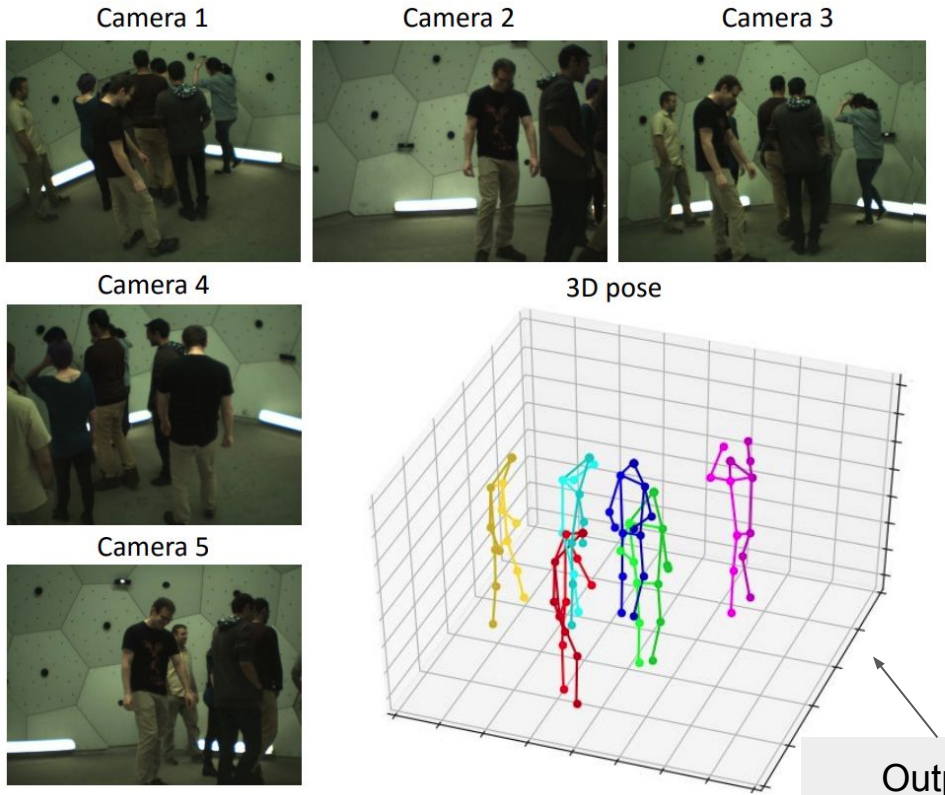


Images credits:
- Rhodin, Helge, et al. "Learning monocular 3d human pose estimation from multi-view images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
- https://becominghuman.ai/what-is-the-main-purpose-of-video-annotation-in-machine-learning-and-ai-11805710bd95

Problem Statement

Inputs

- Multiple calibrated camera views



Our focus
3D reconstruction of a *single* time instant

Output

- 3D skeletons

Motivation

Detections can be noisy, due to (self-)occlusions or uncommon views.
These errors are hard to recover in later stages.

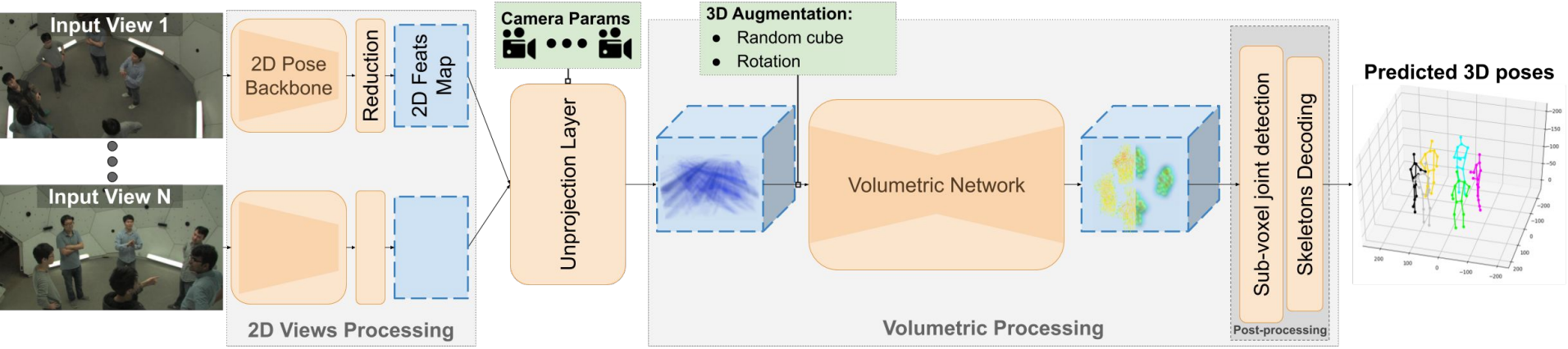
Images credits:
- Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

Motivation

Re-identification techniques have been explored.
They are usually built on top of 2D pose estimation networks.

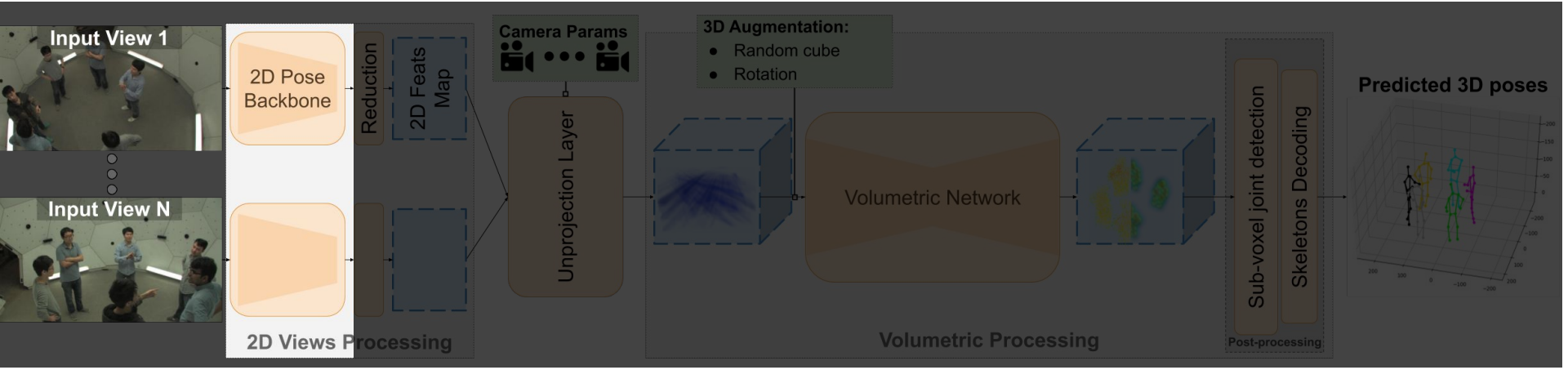
Images credits:
- Xiao, Qiqi, et al. "Cross domain knowledge transfer for person re-identification." arXiv preprint arXiv:1611.06026 (2016).

Architecture



Directly find 3D people poses from multiple calibrated camera views

2D Backbone

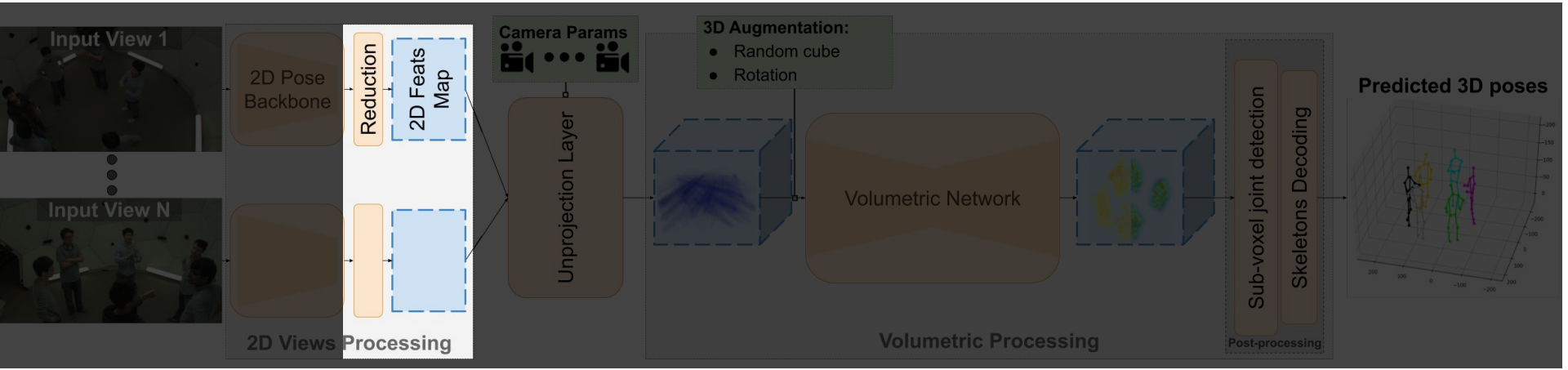


- **Input:** 2D image from single view
- **Output:** 2D features map (512 channels)

- Very fast MobileNet V1 Pretrained on COCO [1]

Lightweight OpenPose [1] references:
https://github.com/Daniil-Osokin/lightweight-human-pose-estimation.pytorch

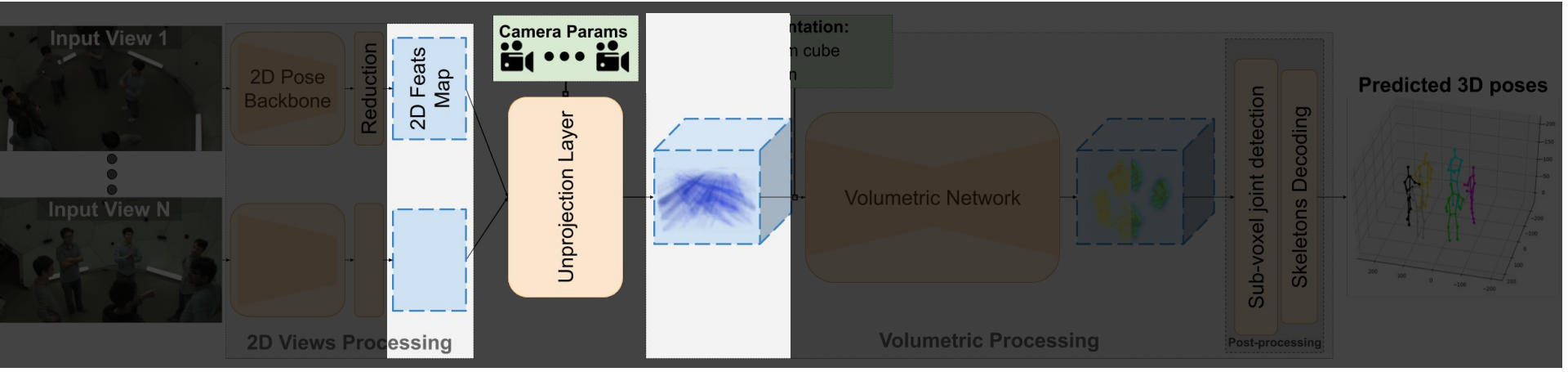
Reduction



- **Input:** 2D features map (512 channels)
- **Output:** 2D features map (64 channels)

- Residual layer with 3 depthwise-separable convs

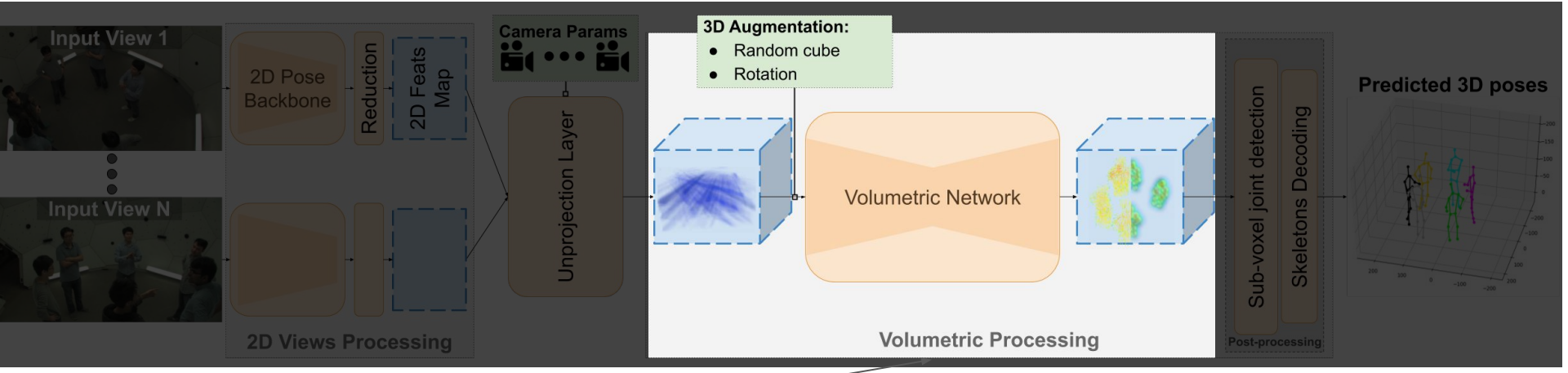
Unprojection Layer



- **Input:**
  - 2D features maps - for every view
  - Camera parameters *(Both Intrinsics and extrinsics)*
- **Output:**
  - A single 3D features cube representing the whole scene

- Not learned
- Very fast implementation in GPU (pytorch)
- Lookup table with interpolation
- Differentiable wrt camera params

Volumetric Network

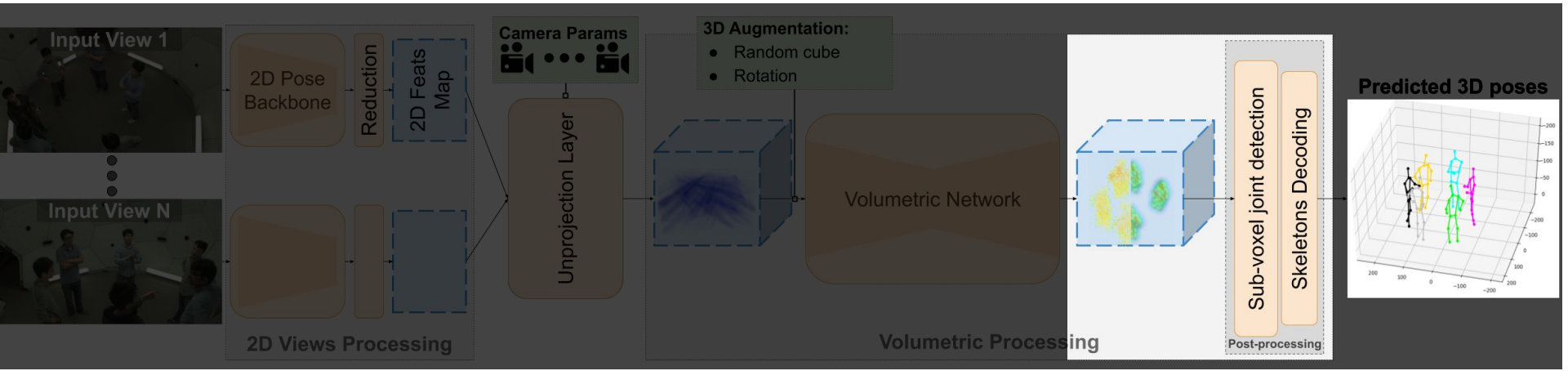


- **Input:**
  - 3D input features
- **Output:**
  - 3D representation of the whole scene. Keypoints + PAFs

- Slightly modified V2V

V2V
https://arxiv.org/pdf/1711.07399v3.pdf

Decoding



- **Input:** 3D heatmap + vectormap
- **Output:** List of 3D poses

- Sub-pixel joint detection
- Skeletons decoding

Datasets

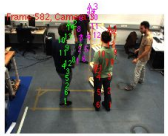
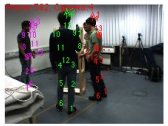
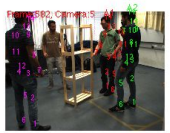
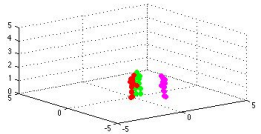
CMU Panoptic [1]

- 30+ HD views
- Hardware-based sync
- Calibration
- 65 sequences (5.5 hours)
- 1.5 millions of 3D skeletons

Shelf [2]

- used to evaluate cross-dataset model generalization
- single scene of four people
- video streams from five calibrated cameras.



H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, “Panoptic studio: A massively multiview system for social motion capture,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342.

Evaluation Metrics

MPJPE: Mean Per Joint Precision Error

Average of the square distance of the predicted joints from the corresponding ground-truth joints

PCP: Percentage of Correctly estimated Parts

a. Implemented according to [1].
b. A body part is correct if the average distance of the two joints is less than a threshold from the corresponding groundtruth joints locations.
c. The threshold is 50% the length of the groundtruth body part.

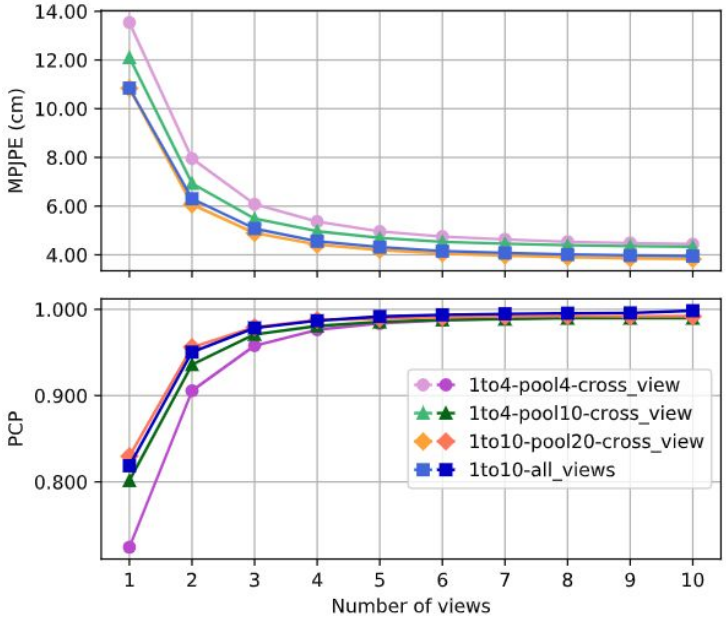
J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou, “Fast and robust multi-person 3d pose estimation from multiple views,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7792–7801.

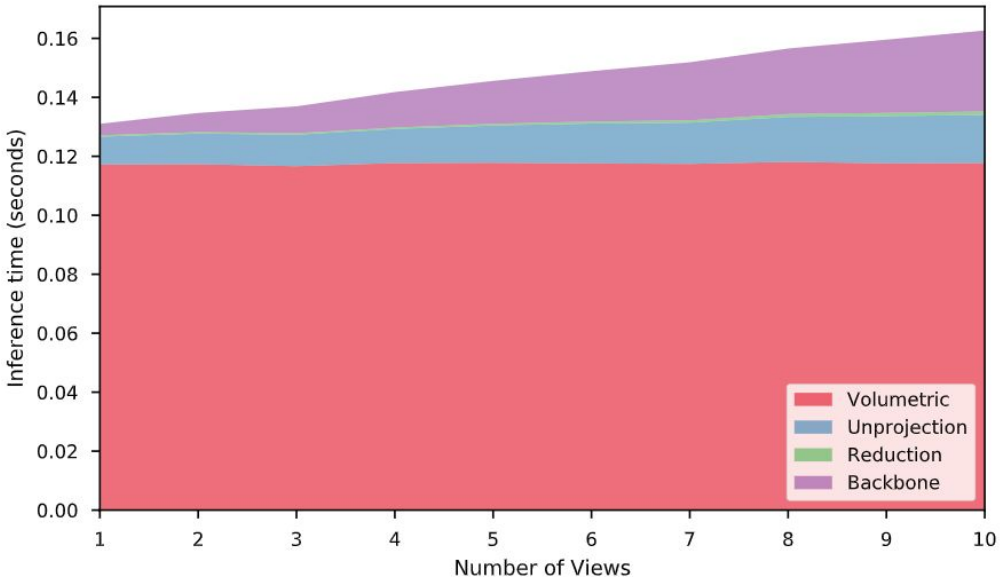
Ablation studies

- 3D Augmentations
- Number of volumetric features
- Loss type
  - Different weights on the heatmap / vectormap loss components
- Sub-voxel refinement

Cube	Rotation	MPJPE (cm)	PCP Head	Torso	Up Arm	Lo Arm	Up Leg	Lo Leg	Avg
3D Augmentations									
		8.236	99.1	99.3	87.8	65.4	96.9	88.3	89.2
✓		4.598	99.6	**99.7**	98.5	90.1	**99.3**	98.5	97.7
	✓	5.350	99.6	**99.7**	98.6	91.1	99.0	94.9	97.3
✓	✓	**3.859**	**99.7**	**99.7**	**99.5**	**95.6**	**99.3**	**98.8**	**98.8**
Number of Volumetric Features									
32		4.760	99.6	99.7	97.1	78.9	99.5	98.6	95.9
64		**3.859**	**99.7**	**99.7**	**99.5**	95.6	**99.3**	**98.8**	98.8
96		3.975	**99.7**	**99.7**	**99.5**	**96.2**	**99.3**	98.7	**98.9**
Loss Type									
L1		4.106	99.6	**99.7**	99.2	96.2	99.0	98.0	98.7
L2		4.125	99.6	**99.7**	**99.5**	**96.6**	**99.4**	**98.9**	**99.0**
SmoothL1		**3.859**	**99.7**	**99.7**	**99.5**	95.6	99.3	98.8	98.8
Heatmap / Vectormap Loss Ratio									
1		**3.859**	**99.7**	**99.7**	**99.5**	95.6	99.3	**98.8**	98.8
3		4.074	**99.7**	**99.7**	99.1	**96.6**	**99.5**	98.6	**98.9**
10		3.935	**99.7**	**99.7**	98.0	90.9	**99.5**	**98.8**	97.9
Sub-voxel refinement									
		4.899	**99.7**	**99.7**	99.4	94.9	**99.3**	**98.8**	98.6
✓		**3.859**	**99.7**	**99.7**	**99.5**	**95.6**	**99.3**	**98.8**	**98.8**

Study on the number of input views





- Assessing cross-view generalization
- Good results even with 1 view

Inference time vs Number of views
Sub-modules decomposition

Assessing generalization

Panoptic D2D test set:

- Unseen views
  (new cameras, new angles)
- Unseen scenes

Model	MPJPE (cm) single	multi	avg	PCP avg
ACTOR [33] (2 views)*	17.21	50.24	33.72	-
ACTOR (4 views)*	8.19	20.10	14.14	-
ACTOR (10 views)*	6.13	12.21	9.17	-
Oracle [33] (using GT to select cameras)*	**4.24**	9.19	6.71	-
Ours (1 unseen view)	10.34	9.32	9.43	80.8
Ours (2 to 4 unseen views depending on scene)	5.30	**4.09**	**4.22**	98.2
Ours (10 views, from training view pool)	**3.50**	**3.56**	**3.55**	**98.6**

*ACTOR: number in brackets refers to maximum number of views to choose from. Oracle means: best views to triangulate are selected using groundtruth.

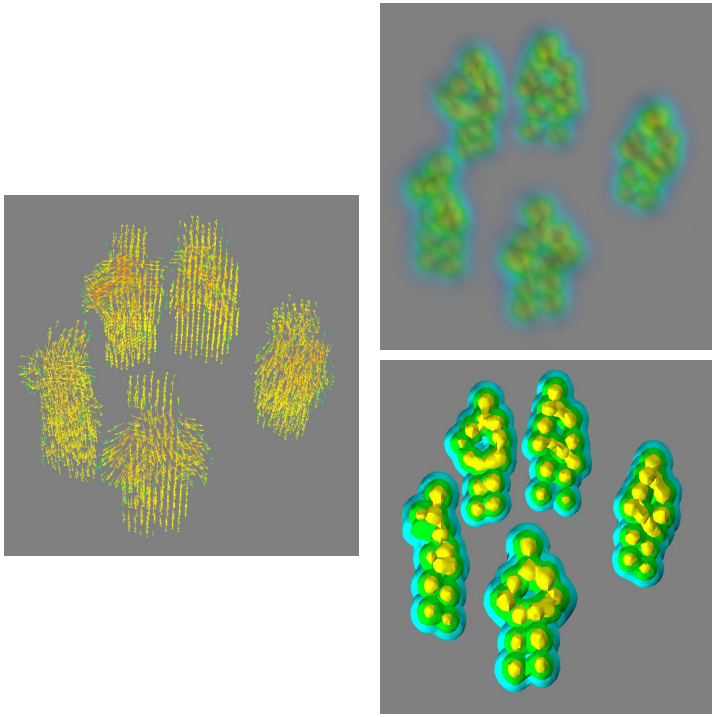
Shelf dataset:

- Completely unseen dataset
- Not yet SOTA results but getting closer
- Probably benefits from variety of pose configurations in training

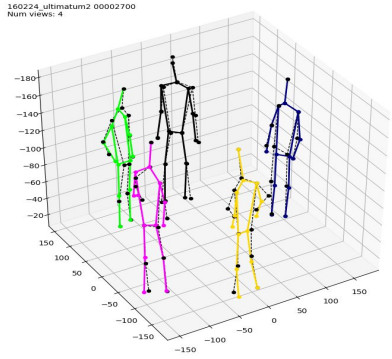
Model	Actor 1	Actor 2	Actor 3	Avg	Speed(s)
Belagiannis et al. [34]	66.1	65.0	83.2	71.4	-
Belagiannis et al. [40]	75.0	67.0	86.0	76.0	-
Belagiannis et al. [41]	75.3	69.7	87.6	77.5	-
Ershadi et al. [42]	93.3	75.9	94.8	88.0	-
Dong et al. [36]	**98.8**	**94.1**	**97.8**	**96.9**	.465
Ours	94.3	78.4	96.8	89.8	**.146**

Visual Results - Multiple views

Input views

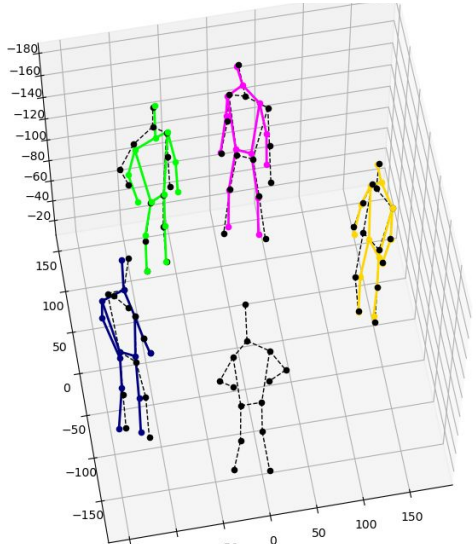


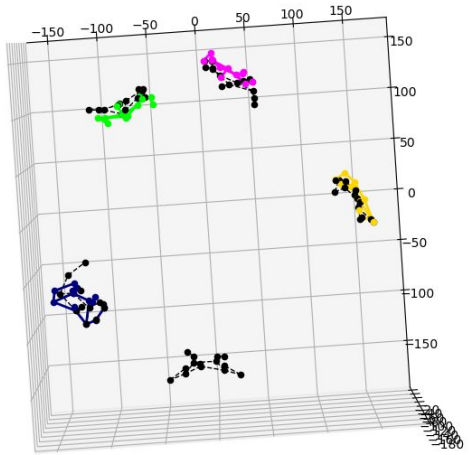
Intermediate representations

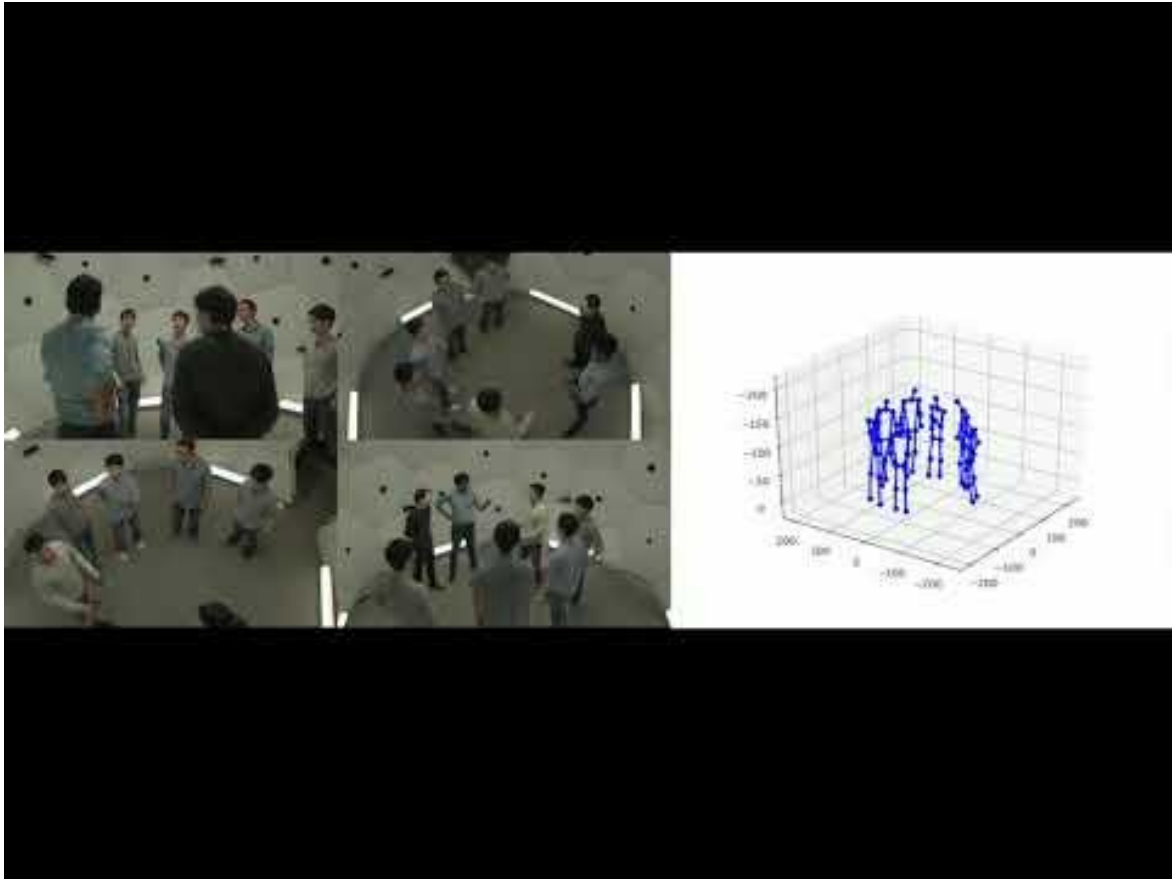


Output

Visual Results - Single View

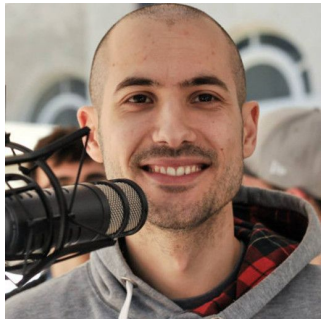






- Unseen scene
- Unseen camera views

Thank you for your attention



Alessio Elmi

Davide Mazzini

Pietro Tortella

{alessio, davide, pietro}@standard.ai

STANDARD

Checkout Technologies
in-store intelligence