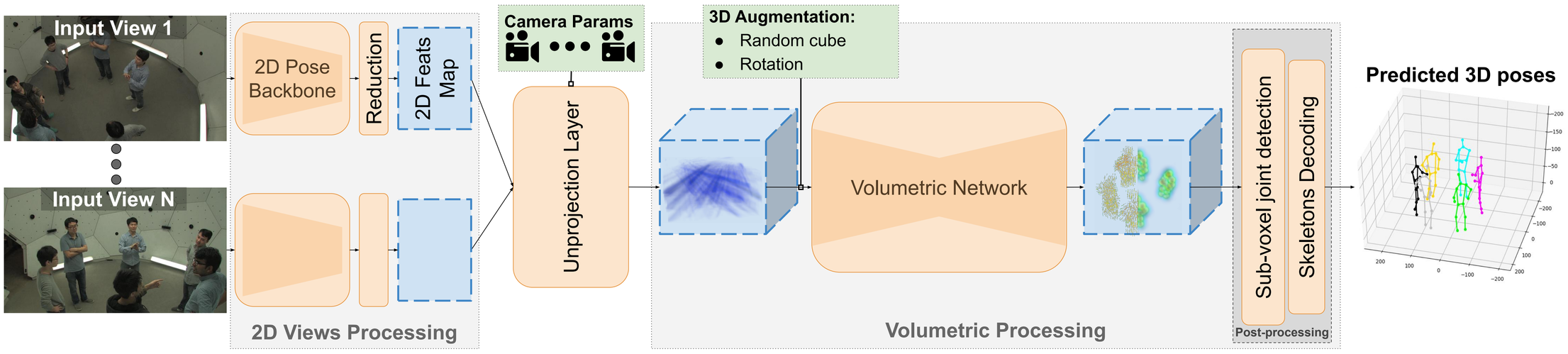
Light3DPose

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

Alessio Elmi - Davide Mazzini - Pietro Tortella
Standard Cognition - Checkout Technologies s.r.l.
{alessio,davide,pietro}@standard.ai

STANDARD
Checkout Technologies
in-store intelligence



Introduction

We propose Light3DPose, a complete bottom-up approach to reconstruct 3D human poses from few calibrated camera views.
With quantitative and qualitative analysis we show that:

- it can handle crowded scenes with good accuracy.
- it scales efficiently with the number of input sources (camera views)
- it is able to produce good results even from a monocular view.

Method

Light3DPose is composed of four main stages:

- A 2D Views Processing stage which returns a 2D feature map for each camera
- An Unprojection layer [6] which aggregates the information coming from all the 2D views into a 3D input features space representation
- A volumetric Processing that process the aggregated 3D representation and produces an intermediate output representation
- A Sub-voxel joint detection and a skeleton decoding part to detect and build complete body skeletons

Datasets and Metrics

We trained and tested our model on CMU Panoptic dataset. We also tested the model on the Shelf [1] in order to evaluate the cross-dataset model generalization.
Metrics employed:

- Mean Per Joint Position Estimation (MPJPE)
- Percentage of correctly estimated Parts (PCP)

Ablation Studies

We performed ablation studies to evaluate the effect on the model's performance with respect to:

- Augmentation strategies
- Number of volumetric features
- Loss function type
- Weighting of the different 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**

Considerations:

- Our synthetic 3D-data augmentation policies greatly enhance the network performance
- Our sub-voxel localization strategy overcomes the issue of a coarse quantized 3D space thus substantially impacting the MPJPE localization measure.

Quantitative results

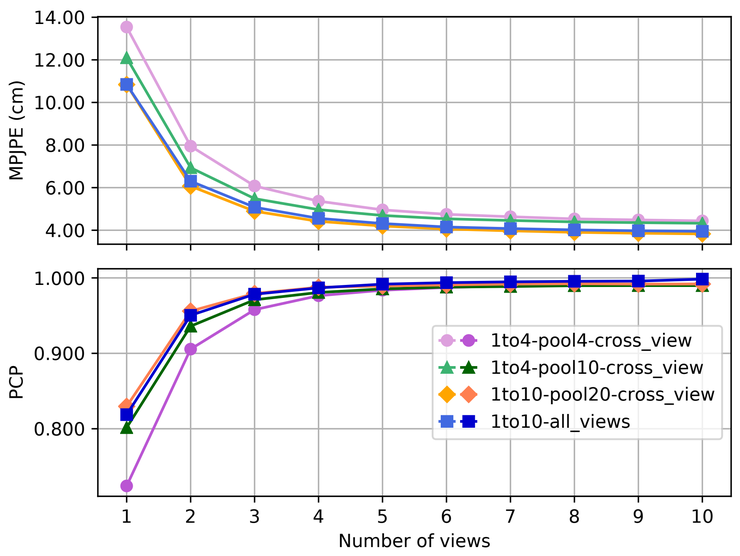
Panoptic

We compare our results on Panoptic dataset with [7] on different train/test regimes:

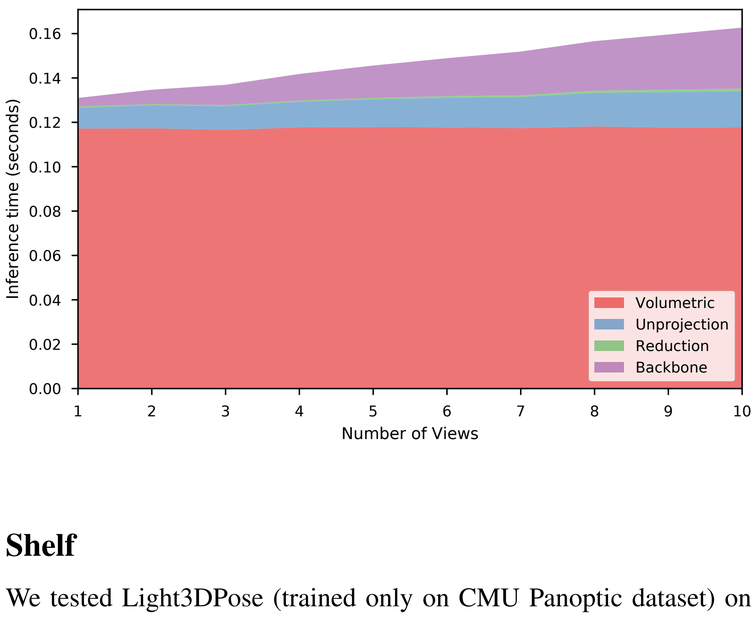
Model	MPJPE (cm) single	multi	avg	PCP avg
ACTOR [7] (2 views)*	17.21	50.24	33.72	-
ACTOR (4 views)*	8.19	20.10	14.14	-
ACTOR (10 views)*	6.13	12.21	9.17	-
Oracle [7] (using GT to select cameras)*	**4.24**	9.19	6.71	-
Ours (1 unseen view)	10.34	9.32	9.43	80.8
Ours (2 to 4 unseen views depending on scene)	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.

Light3DPose can even process a scene from a single view. The accuracy increases adding more views to it.



A break down of the inference time for each single component. Since the view-dependent part is very lightweight, the inference time scales well with an increasing number of input views.



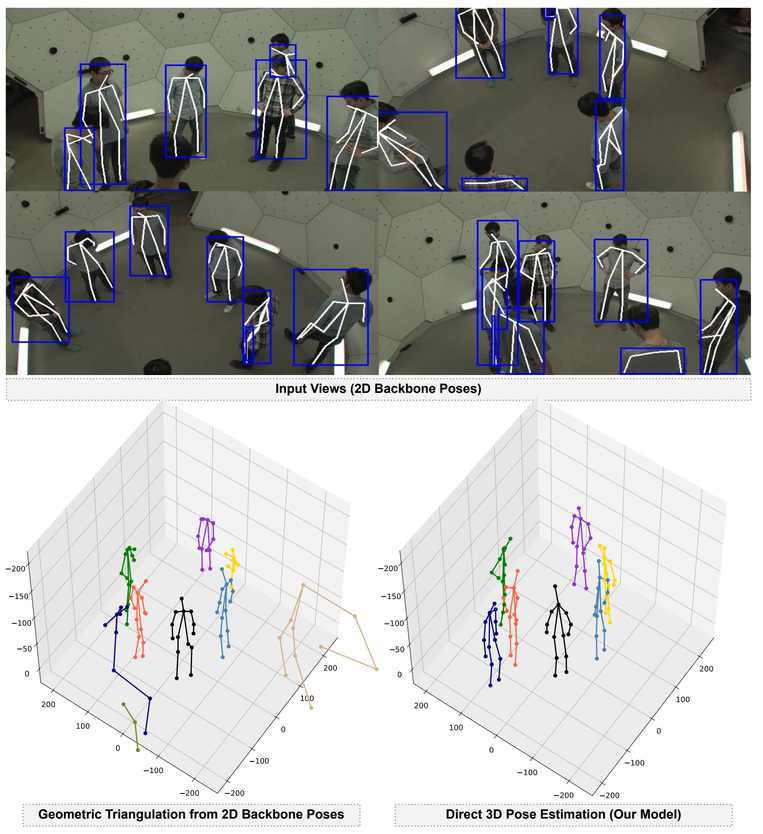
Shelf

We tested Light3DPose (trained only on CMU Panoptic dataset) on Shelf [1], and achieve SOTA-comparable results using only one third of the computational time.

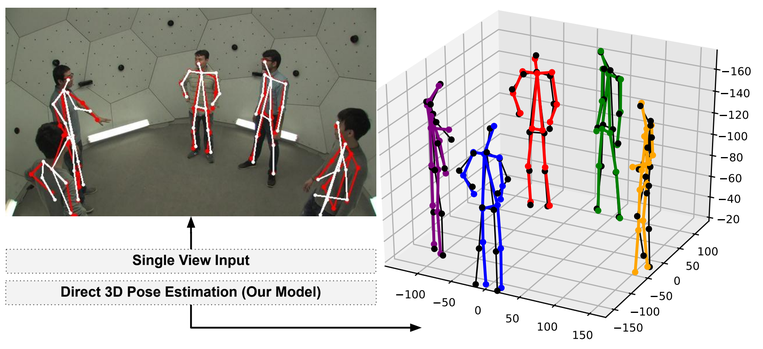
Model	Actor 1	Actor 2	Actor 3	Avg	Speed(s)
Belagiannis et al. [1]	66.1	65.0	83.2	71.4	-
Belagiannis et al. [3]	75.0	67.0	86.0	76.0	-
Belagiannis et al. [2]	75.3	69.7	87.6	77.5	-
Ershadi et al. [5]	93.3	75.9	94.8	88.0	-
Dong et al. [4]	**98.8**	**94.1**	**97.8**	**96.9**	.465
Ours	94.3	78.4	96.8	89.8	**.146**

Visual Results

The power of direct 3D pose estimation: by exploiting a holistic 3D space representation, our volumetric architecture can learn strong pose priors and implicitly discards false detections, being less prone to occlusion-related errors and better dealing with crowded scenes.



An example of Light3DPose results with a single input view.



Conclusions

- The proposed method achieves SOTA accuracy on Panoptic dataset and comparable accuracy on Shelf (never seen in training and validation);
- it is 3x faster than SOTA 3D Pose estimation algorithms based on 2D Pose detection + Triangulation
- A larger dataset with higher variety of poses will probably lead to further improvements.

References

[1] B. et al. 3d pictorial structures for multiple human pose estimation.

[2] B. et al. 3d pictorial structures revisited: Multiple human pose estimation.

[3] B. et al. Multiple human pose estimation with temporally consistent 3d pictorial structures.

[4] D. et al. Fast and robust multi-person 3d pose estimation from multiple views.

[5] E.-N. et al. Multiple human 3d pose estimation from multiview images.

[6] I. et al. Learnable triangulation of human pose.

[7] P. et al. Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction.