Supplementary Material

Matteo Fabbri[1*] Fabio Lanzi[1] Stefano Alletto[2] Simone Calderara[1] Rita Cucchiara[1]
[1]University of Modena and Reggio Emilia
{name.surname}@unimore.it
[2]Panasonic R&D Company of America
{name.surname}@us.panasonic.com

1. Implementation Details

For the sake of reproducibility, in this section we illustrate the architectures of the Code Predictor and the Pose Refiner modules of our LoCO pipeline.

Code Predictor Inspired by [4], our method simply adds a few convolutional layers (f-c2d) to the last convolution stage of a backbone network. Tab. 1 reports the detailed structures of the various f-c2d blocks utilized in our experiments. *ConvTr2D* refers to transposed 2D convolutions while *Conv2D* refers to simple 2D convolutions. For each layer, we provide: number of input channels, number of output channels, kernel size and stride. In all the proposed experiments we utilized InceptionV3 [3] pretrained on Imagenet [1] as backbone architecture.

Pose Refiner The structure of the Pose Refiner is shown in Fig. 1. It is a simple network composed by three fully connected layers with ReLU activation followed by a skip connection. Input and output are normalized root-relative representations of a single 3D pose, with values in range $[0, 1]$. During training, Gaussian noise (mean: $0m$, variance: $0.08m$) is applied to the input pose while some joints are randomly removed with probability 0.1. The removed joints are coded with a default value of $(-1, -1, -1)$.

2. Volumetric Heatmap Spaces

In our experiments, we defined our Volumetric Heatmap representation according to two different pseudo-3D spaces, depending on which dataset we used:

- The first space is defined as $S_1 = D \times H' \times W'$, where H' and W' are the height and width, downsampled by a actor of 8, of the image plane and D is the maximum distance from the camera in meters, quantized with 316 bins. We adopted S_1 for JTA.

* Work done while interning at Panasonic R&D Company of America

	layer	in ch.	out ch.	ker.	str.
LoCO$^{(1)}$	ConvTr2D→ReLU	F	1024	4	2
	ConvTr2D→ReLU	1024	512	4	2
	Conv2D→ReLU	512	256	4	1
	Conv2D	256	158	1	1
LoCO$^{(2)}$	ConvTr2D→ReLU	F	1024	4	2
	Conv2D→ReLU	1024	512	4	1
	Conv2D→ReLU	512	256	4	1
	Conv2D	256	79	1	1
LoCO$^{(3)}$	Conv2D→ReLU	F	1024	4	1
	Conv2D→ReLU	1024	512	4	1
	Conv2D→ReLU	512	256	4	1
	Conv2D	256	39	1	1

Table 1. Structure of the three f-c2d block variants of the Code Predictor used for our HPE experiments. F represents the number of output channels of the exploited feature extractor. In all our experiments we used Inception v3 with $F = 2048$

- The second space is defined as $S_2 = Z \times H' \times W'$, where H' and W' are defined as in S_1, and Z is the maximum $\mathbf{z}$ axis value of the real 3D space in the standard coordinate system centered to the camera. Z is expressed in meters and quantized with 316 bins. We adopted S_2 for Panoptic and Human3.6m.

Although the difference between these two spaces is minimal, we adopted S_1 for JTA because this dataset already provide a maximum camera distance, which is 100 meters.

3. Detection Experiments

To show how our LoCO approach can be effectively adopted also for the detection task in crowded scenarios under heavy occlusions, we have tested our system in terms of 2D people detection comparing it with YOLOv3 [2] on the JTA test set. Using LoCO, we predict 3D poses and project them on 2D bounding boxes using the camera intrinsic parameters.

In terms of precision, recall, and F1 (with the bounding

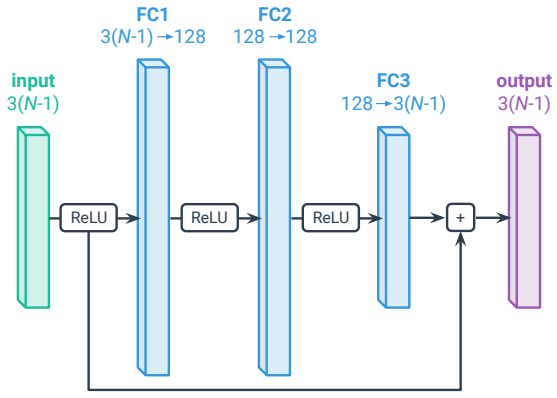

Figure 1. Structure of our Pose Refiner: 3 fully connected layers with ReLU activation followed by a skip connection. N is the number of joints. In all our experiments we considered $N = 14$.

box IoU threshold at 0.5), using our LoCO$^{(2)}$+ trained on JTA, we get 81.94, 69.73, and 75.39 respectively; with out of the box YOLOv3, instead, we obtain 99.12, 30.81 and 44.50.

Although our model is less precise than YOLOv3 (around -20%), it surpasses it by a large margin (around +40%) in terms of recall, resulting in an F1-score that is clearly in our favor (almost +30%). The scenes in JTA, in fact, are extremely crowded and present a very high percentage of occlusion with multiple overlapping people. It is not easy for a detector to handle situations of this type, while a part-based bottom-up method is much less affected by this problem.

4. Skeleton Grouping Details

Let's consider K different type of joints and $N_0, \dots, N_{K-1}$ number of detections for each joint type. Given N_0 predicted heads, $\mathbf{j}_{0,0}, \dots, \mathbf{j}_{0,N_0-1} \in \mathbb{R}^3$, and $N_k, k \in [1, K-1]$ predicted joints of another type, $\mathbf{j}_{k,0}, \dots, \mathbf{j}_{k,N_k-1} \in \mathbb{R}^3$, we define $K-1$ cost matrices, $\mathfrak{D}_1, \dots, \mathfrak{D}_{k-1}$, as follows: $\mathfrak{D}_k : \{0, \dots, N_0 - 1\} \times \{0, \dots, N_k - 1\} \to \mathbb{R}$ where each element $d_{a,b}$ is defined as

$$d_{a,b} = \begin{cases} \|j_{0,a} - j_{k,b}\| & \text{if } \|j_{0,a} - j_{k,b}\| \leq 1.5 \cdot \tau_k \\ +\infty & \text{otherwise} \end{cases} \quad (1)$$

The threshold τ_k in (1) is the maximum distance between a head and a joint of type k (belonging to the same person) on the entire training set. For each $k = 1, \dots, K-1$, joint-head associations are made with the Hungarian algorithm using $\mathfrak{D}_k$ as cost matrix. The same joint-grouping procedure is applied on both multi-person datasets. By removing the anatomical constraints, results on Panoptic show an MPJPE degradation of about 9 millimeters while on JTA, no degradation in terms of metric has been observed.

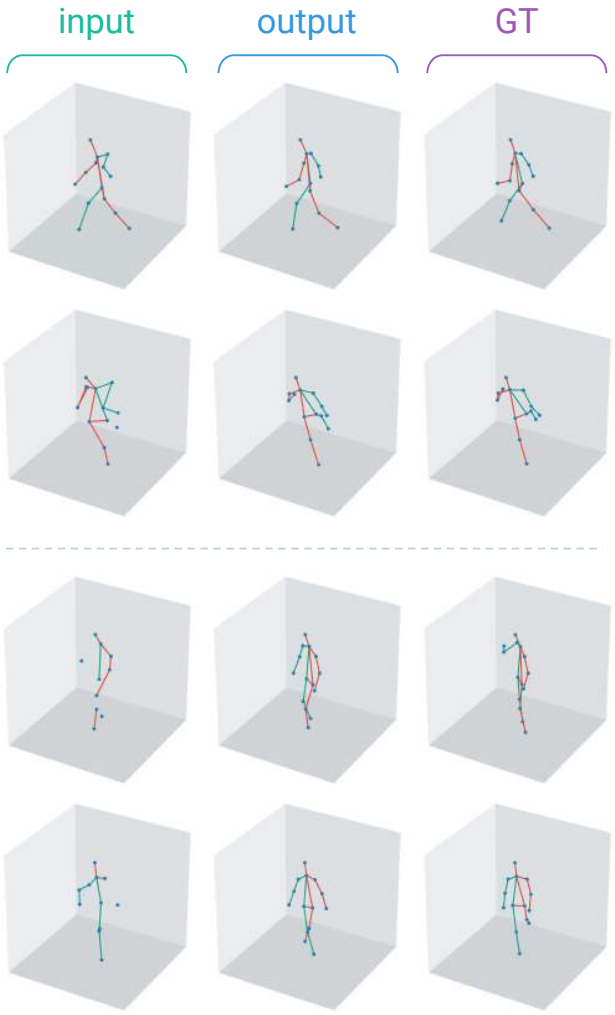

Figure 2. Qualitative results of our Pose Refiner model on the JTA dataset. 1st and 2nd rows: examples where the output is anatomically plausible and consistent with the ground truth; 3rd and 4th rows: examples where the output is anatomically plausible, but inconsistent with the ground truth.

5. Additional Qualitative Results

We report the results of our Pose Refiner on JTA. When the input pose fed to our Pose Refiner has few missing joints and position errors, the reconstruction is consistent with the ground truth pose (Fig. 2 - first two rows). Conversely, when the input pose is more degraded, the reconstruction is still plausible, but not always coherent with the ground truth (Fig. 2 - last two rows). Additionally, Fig. 3 depicts some qualitative results of LoCO on JTA, CMU Panoptic and Human3.6m. Our method can be applied to both crowded scenarios and single person contexts, displaying good generalization capabilities in a wide range of contexts.

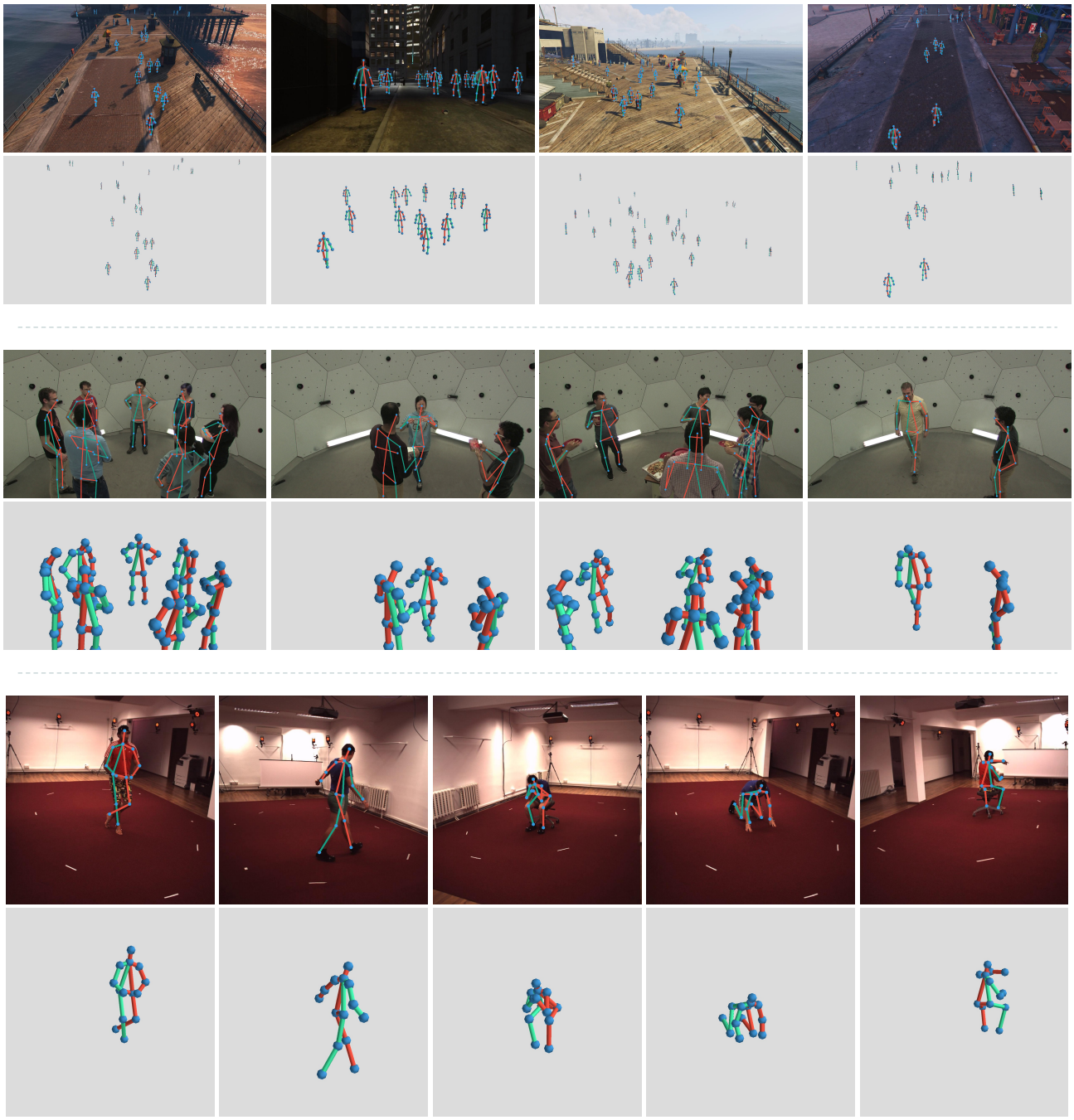

Figure 3. Additional qualitative results of our LoCO approach. 1st and 2nd rows: result of LoCO$^{(2)}+$ on the JTA dataset; 3rd and 4th rows: result of LoCO$^{(2)}+$ on the CMU Panoptic dataset; 5th and 6th rows: result of LoCO$^{(3)}+$ on the Human3.6m dataset

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE CVPR*, 2009. 1

[2] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE CVPR*, 2016. 1

[4] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on Computer Vision (ECCV)*, 2018. 1