Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation Supplementary Material

Bugra Tekin Pablo Márquez-Neila Mathieu Salzmann Pascal Fua
CVLab, EPFL, Switzerland
{firstname.lastname}@epfl.ch

In this supplementary material, we analyze the influence of our regularization term encouraging sharp fusion in Eq. 4, provide running time for our algorithm, and show additional qualitative results on the Human3.6m [3], HumanEva-I [6], KTH Multiview Football II [1] and Leeds Sports Pose [4] datasets.

Effect of the regularization. Below, we analyze the effect of the regularization term that encourages sharp fusion in Eq. 4. In the absence of the regularization term, the network mixes the data and fusion streams without necessarily fusing them at a specific layer. As discussed in the main paper, this corresponds to a model with many active parameters. Therefore it is prone to overfitting and computationally less efficient at test-time. In Table 1, we compare the results of our approach with and without this regularization term. For the latter, we do not parametrize the weights of the network with a sigmoid function and do not constrain the network to have a sharp fusion. The results confirm that encouraging sharp fusion yields both better accuracy and faster prediction.

Method	3D Pose Error	Runtime
Without regularization	68.30	0.013
With regularization	60.17	0.006

Table 1. Quantitative results of our fusion approach with and without the regularization term encouraging sharp fusion. These experiments were carried out on the *Eating* action class of Human3.6m. 3D pose error is computed as the average Euclidean distance (in milimeters) between the predicted and ground-truth 3D joint positions. Runtime denotes the computational time spent, in sec/frame, during testing for the fusion network with and without the regularization term. With the regularization term, inactive layers are pruned after training, which yields a more efficient network for test-time prediction.

Running time. We carried out our experiments on a machine equipped with an Intel Xeon CPU E5-2680 and an NVIDIA TITAN X Pascal GPU. It takes 90 ms to compute 2D joint location confidence maps and 6 ms to predict 3D pose with our fusion network. Therefore, the total runtime of our method is 0.096 sec/frame (over 10 fps), which compares favorably with the recent model-based methods ranging from 0.04 fps to 1 fps [7, 5, 8].

Additional qualitative results. We provide additional qualitative results for the KTH Multiview Football II [1], Human3.6m [3] and HumanEva [6] datasets in Figs. 1, 2 and 3, respectively. Finally, we demonstrate that our regressor trained on the recently released synthetic dataset of [2] generalizes well to real images obtained from the Leeds Sports Pose dataset [4] in Fig. 4. Additional qualitative results can be found in the accompanying videos.

References

[1] M. Burenius, J. Sullivan, and S. Carlsson. 3D Pictorial Structures for Multiple View Articulated Pose Estimation. In *CVPR*, 2013.

[2] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-or, and B. Chen. Synthesizing Training Images for Boosting Human 3D Pose Estimation. In *3DV*, 2016.

[3] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *PAMI*, 2014.

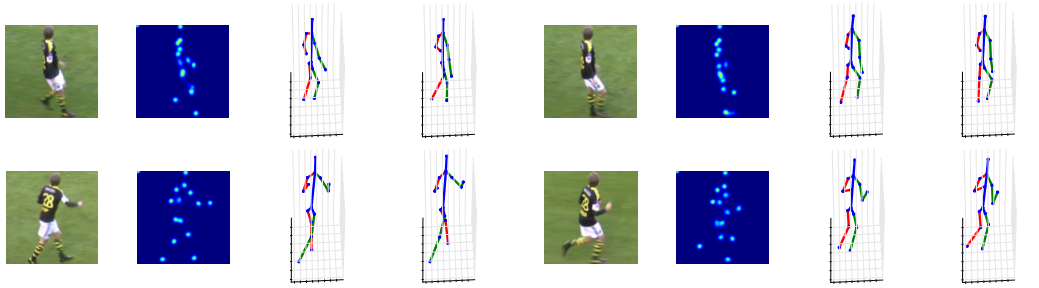
[4] S. Johnson and M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*, 2010.

[5] M. Sanzari, V. Ntouskos, and F. Pirri. Bayesian Image Based 3D Pose Estimation. In *ECCV*, 2016.

[6] L. Sigal and M. Black. Humaneva: Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion. Technical report, Department of Computer Science, Brown University, 2006.

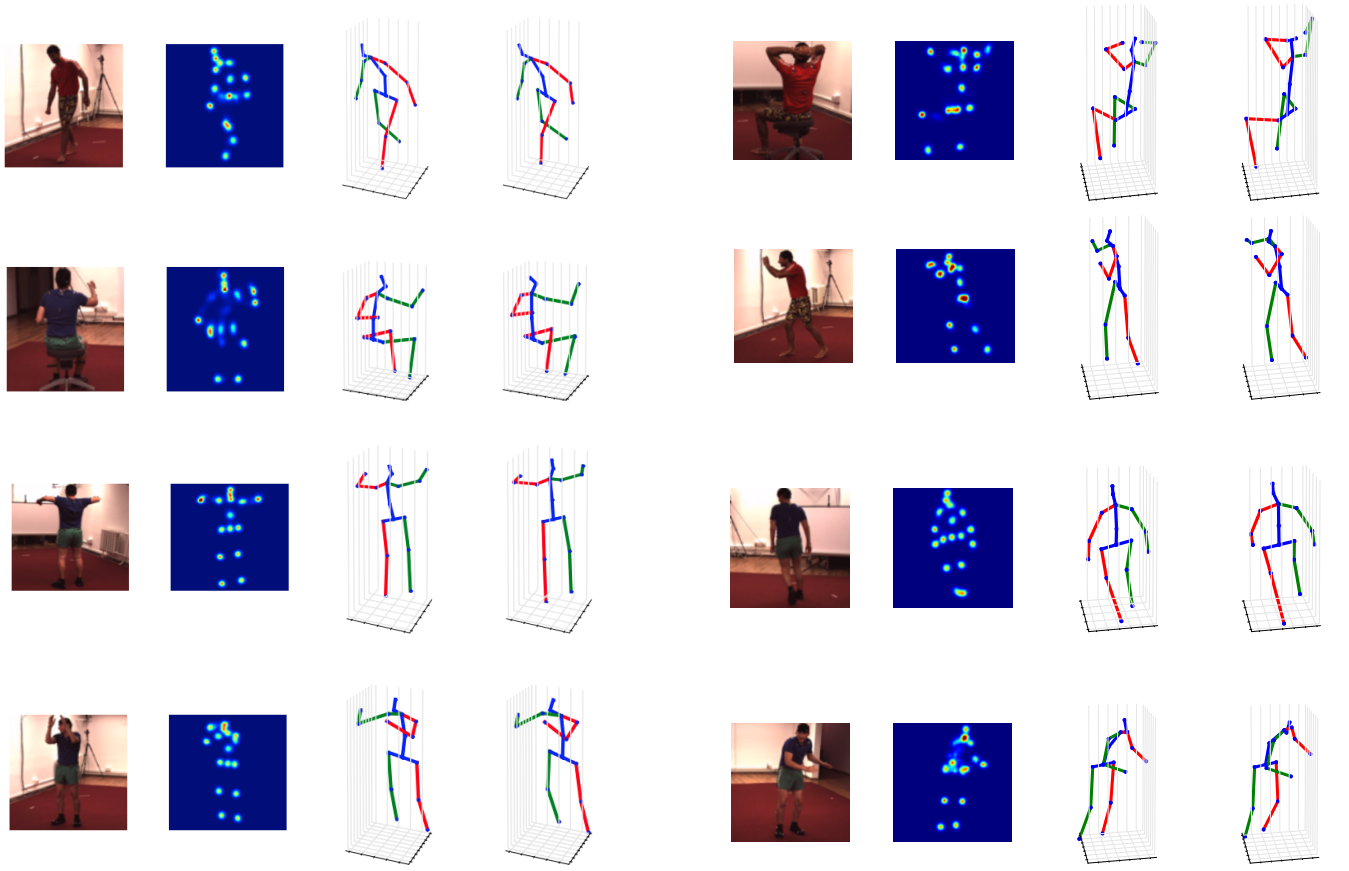
[7] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A Dual-Source Approach for 3D Pose Estimation from a Single Image. In *CVPR*, 2016.

[8] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *CVPR*, 2016.



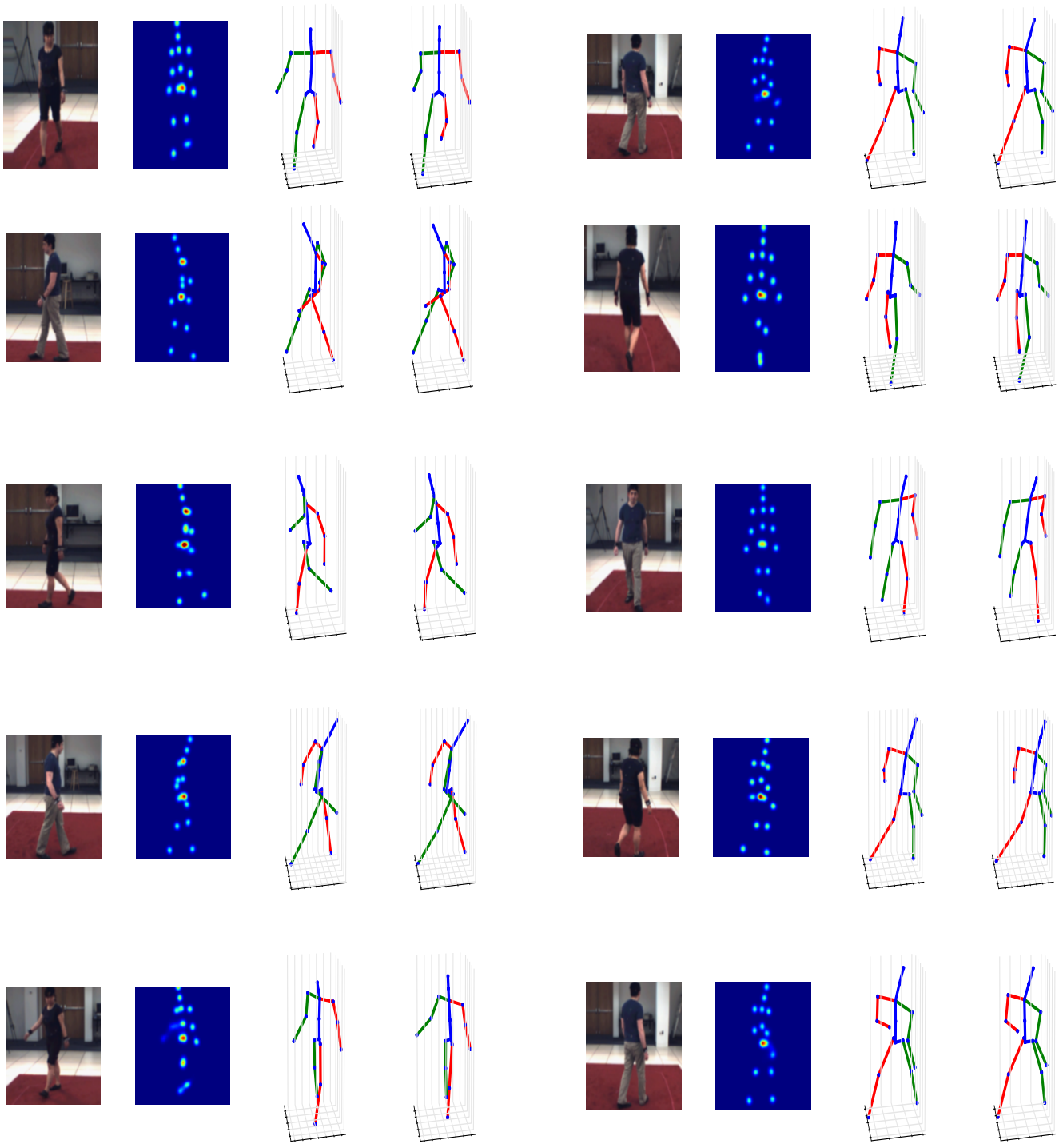
(a) Image (b) Confidence Map (c) Prediction (d) Ground-truth (e) Image (f) Confidence Map (g) Prediction (h) Ground-truth

Figure 1. Pose estimation results on KTH Multiview Football II. **(a, e)** Input images. **(b, f)** 2D joint location confidence maps. **(c, g)** Recovered pose. **(d, h)** Ground truth. Best viewed in color.



(a) Image (b) Confidence Map (c) Prediction (d) Ground-truth (e) Image (f) Confidence Map (g) Prediction (h) Ground-truth

Figure 2. Pose estimation results on Human3.6m. **(a, e)** Input images. **(b, f)** 2D joint location confidence maps. **(c, g)** Recovered pose. **(d, h)** Ground truth. Note that our method can recover the 3D pose in these challenging scenarios, which involve significant amounts of self occlusion and orientation ambiguity. Best viewed in color.



(a) Image (b) Confidence Map (c) Prediction (d) Ground-truth (e) Image (f) Confidence Map (g) Prediction (h) Ground-truth

Figure 3. Pose estimation results on HumanEva-I. **(a, e)** Input images. **(b, f)** 2D joint location confidence maps. **(c, g)** Recovered pose. **(d, h)** Ground truth. Best viewed in color.

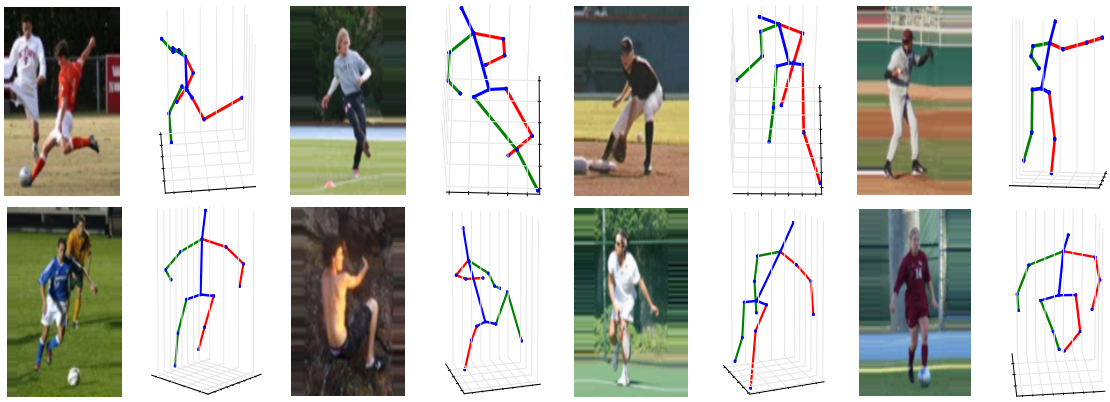


Figure 4. Pose estimation results on LSP. We trained our network on the recently released synthetic dataset of [2] and tested it on the LSP dataset. The quality of the 3D pose predictions demonstrates the generalization of our method. Best viewed in color.