Supplementary Material
Weakly-Supervised 3D Human Pose Learning via Multi-view Images in the Wild

Umar Iqbal Pavlo Molchanov Jan Kautz
NVIDIA
{uiqbal, pmolchanov, jkautz}@nvidia.com

We provide implementation details to reproduce the results in the paper.

1. Implementation Details

We adopt HRNet-w32 [1] as the back-bone of our network architecture. We pre-train the model for 2D pose estimation before introducing weakly-supervised losses. This ensures that the 2D pose estimates are sufficiently good to enforce multi-view consistency $\mathcal{L}_{\text{MC}}$. We use MPII dataset for pre-training. The additional weights for latent depth-maps are not pre-trained. We use a maximum of four camera views C_n=4 to calculate $\mathcal{L}_{\text{MC}}$. If a sample contains more than four views, we randomly sample four views from it in each epoch. We train the model with a batch size of 256, where each batch consists of 128 images with 2D pose annotations and 32 unlabeled multi-view samples (32×4=128 images). For pre-processing, we use a person bounding-box to crop the person into a 256×256 image such that the person is centered and covers roughly 75% of the image. The training data is augmented by random scaling ($\pm 20\%$) and rotation ($\pm 30\%$ degrees). We found that the training converges after 60k iterations. The learning rate is set to $5\text{e}{-4}$, which drops to $5\text{e}{-5}$ at 50k iterations following the Adam optimization algorithm. We use λ=50. Since the training objectives (6) and (7) consist of multiple loss terms, we balance their contributions by empirically choosing ψ=5, α=10, and β=100. Since our pose estimation model estimates absolute 3D pose up to a scaling factor, during inference, we approximate the scale using mean bone-lengths from the training data:

$$\hat{s} = \underset{s}{\text{argmin}} \sum_{j,j' \in \mathcal{E}} (s \cdot \|\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_{j'}\| - \mu^L_{j,j'})^2, \qquad (1)$$

where is $\mu^L_{j,j'}$ is the mean length of the limb formed by joint pair (j, j'). In all of our experiments, we use mean lengths from the training set of H36M dataset.

References

[1] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1