Requested Additional Results.

PoseTrack17 Results on Test Set (R2). We evaluate our model on the PoseTrack17 test set (using our spatiotemporal pose aggregation scheme) and obtain 77.94 mAP, ranking first on the PoseTrack17 leaderboard. We will add these results to our final paper. We will also update Table 2 with missing entries from the leaderboard. We will cite the missing papers and add them to our related work discussion.

PoseTrack18 Results (R2). We also evaluate our model on the PoseTrack18 dataset. Our spatiotemporal pose aggregation scheme yields 80.1 and 78.0 mAP on the PoseTrack18 validation and test sets, respectively, ranking first among entries that use only PoseTrack and COCO data, and second overall. We will include these results as well.

Ablation Study on Dilated Convolution (R1, R2, R3). Here we study the effect of different levels of dilated convolutions in our PoseWarper architecture. We evaluate all these variants on the task of video pose propagation. First, we report that removing dilated convolution blocks from the original architecture reduces the accuracy from 88.7 mAP to 87.2 mAP. We also note that a network with a single dilated convolution (using a dilation rate of 3) also yields 87.2 mAP. Adding a second dilated convolution level (using dilation rates of $3, 6$) improves the accuracy to 88.0. Three dilation levels (with dilation rates of $3, 6, 12$) yield a mAP of 88.4 and four levels (dilation rates of $3, 6, 12, 18$) give a mAP of 88.6. A network with 5 dilated convolution levels (Fig. 2 in the original draft) yields 88.7 mAP. Adding more dilated convolutions does not improve the performance further. Additionally, we also experimented with two networks that use dilation rates of $1, 2, 3, 4, 5$, and $4, 8, 16, 24, 32$, and report that such models yield mAPs of 88.6 and 88.5, respectively, which are slightly lower. We will add this ablation study to our final paper.

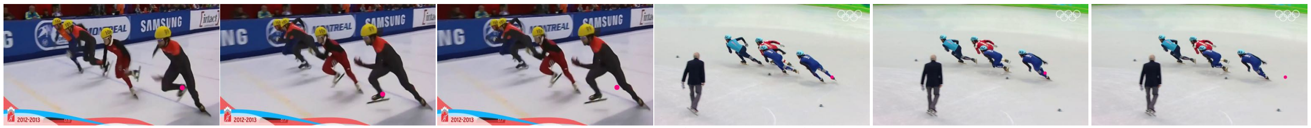
Requested Clarifications.

Spatiotemporal Pose Aggregation. (R1) We chose to average the warped heatmaps from neighboring frames during spatiotemporal pose aggregation, as we discovered it to be a simple yet effective scheme.

Warping Heatmaps (R1). The offsets are used to warp the pose heatmaps. We predict $c \times k_h \times k_w \times 2$ offset channels (i.e., (x, y) displacements) for every pixel where c is the number of joints, and k_h, k_w are the deformable convolution kernel height and width respectively (see L271-273). In our case, $c = 17$, and $k_h = k_w = 3$, which means that we predict 153 (x, y) displacements (306 channels) for every pixel.

Interpretability of Offsets and Comparison with FlowNet2 (R1). We agree with R1 that it is difficult to understand what our predicted offsets encode based on their direct visualizations. However, Figure 5 reveals that different offset maps encode different motions and suggests that the method performs some sort of motion decomposition corresponding to different body parts or discriminative regions. Figure A1 of this document includes a more intuitive illustration of the motion encoded by PoseWarper and compares it to the optical flow computed by FlowNet2 (as requested by R1) for the video pose propagation task. The first frame in each 3-frame sequence illustrates a *labeled* reference frame at time t. For a cleaner visualization, we show only the "right ankle" body joint for one person, which is marked with a **pink** circle in each of the frames (please zoom in). The second frame depicts our propagated "right ankle" detection from the labeled frame in time t to the unlabeled frame in time t+1. The third frame shows the propagated detection in frame t+1 produced by the FlowNet2 baseline. This visualization and similar ones that we generated and that we plan to include in supplementary material, suggest that FlowNet2 struggles to accurately warp poses if 1) there is large motion, 2) occlusions, or 3) blurriness. In contrast, our PoseWarper handles these cases robustly, which is also indicated by our results in Table 1 of our submission (i.e., 88.7 vs 83.8 mAP w.r.t. FlowNet2). For the final version of our paper, we will add more visualizations such as the ones in Figure A1. We believe that they will provide a better qualitative understanding of why our method is advantageous compared to FlowNet2 (besides the quantitative benefits of lower computational cost and improved accuracy, which we already discussed in our submission).

Track ID Annotations (R2). Our approach does *not* require track ID manual annotations or externally generated person tracks. We train our model to warp pose heatmaps from an unlabeled Frame B to a labeled Frame A. For each labeled bounding box of a person in Frame A, we crop Frame B at the same location using a bounding box that is large enough to include the same person even if he/she moved. During spatiotemporal pose aggregation inference, we apply the same scheme and use bounding boxes automatically extracted with a detector from [48] (see L163-165). We note that because our predicted offsets encode motion cues between Frames A and B, we can potentially leverage our offsets for tracking. Furthermore, as noted by R3, our proposed framework is general enough to be applied for other video tasks such as object detection or instance segmentation in video.



Ground Truth Reference Frame (time t) | PoseWarper (time t+1) | FlowNet2 (time t+1) | Ground Truth Reference Frame (time t) | PoseWarper (time t+1) | FlowNet2 (time t+1)

Figure A1: Comparing our PoseWarper and FlowNet2 for video pose propagation. Please zoom in to see the **pink** circle on the right ankle. Unlike our model, FlowNet2 fails to accurately propagate poses when there is large motion, bluriness or occlusions.