Learning Temporal Pose Estimation from Sparsely Labeled Videos

Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, Lorenzo Torresani

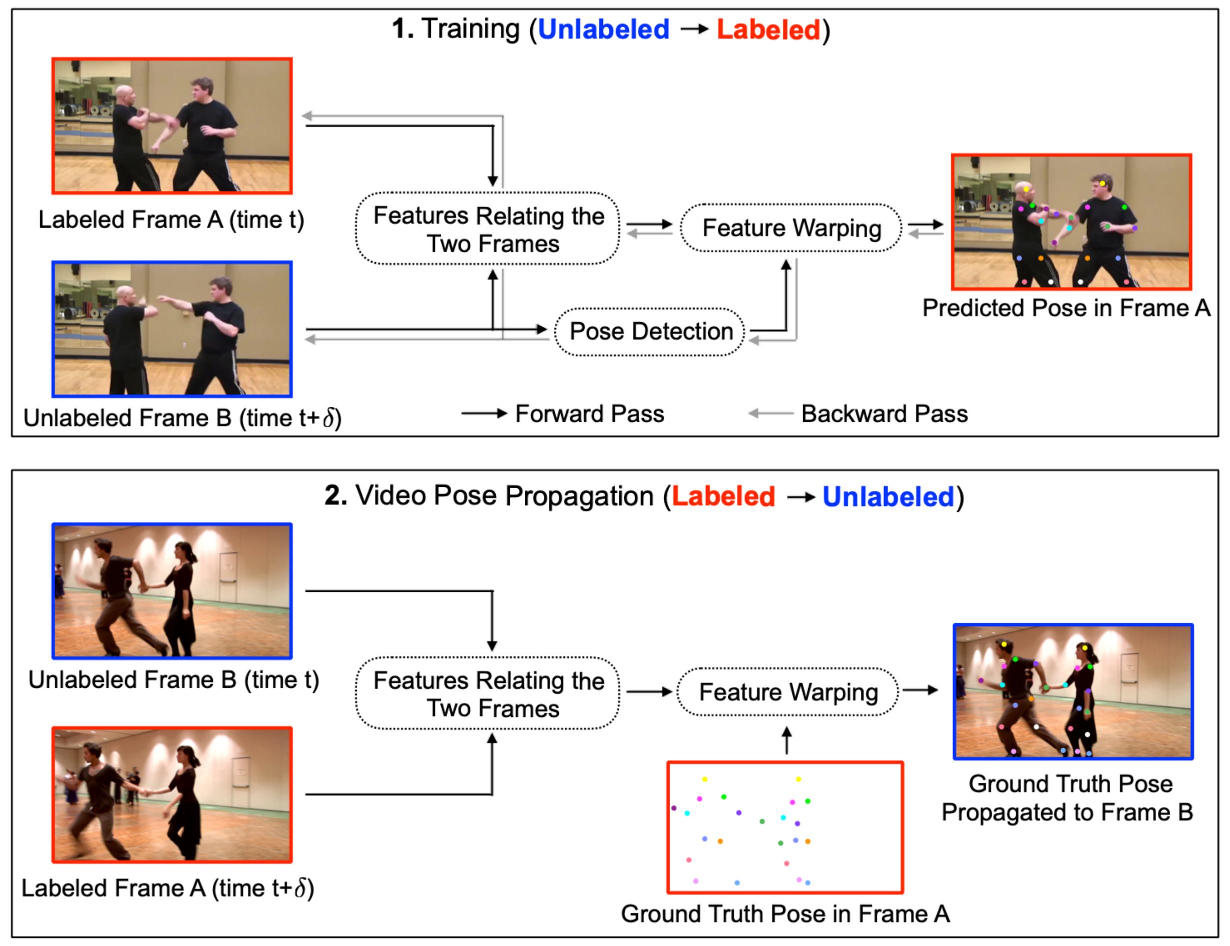
I. Introduction

Problem Overview

- Pose detection in video is challenging due to video defocus, occlusions and, motion blur.
- Densely labeling every frame with multi-person pose annotations is costly and time consuming.
- Videos have high informational redundancy (the content changes little from frame to frame).

High Level Approach

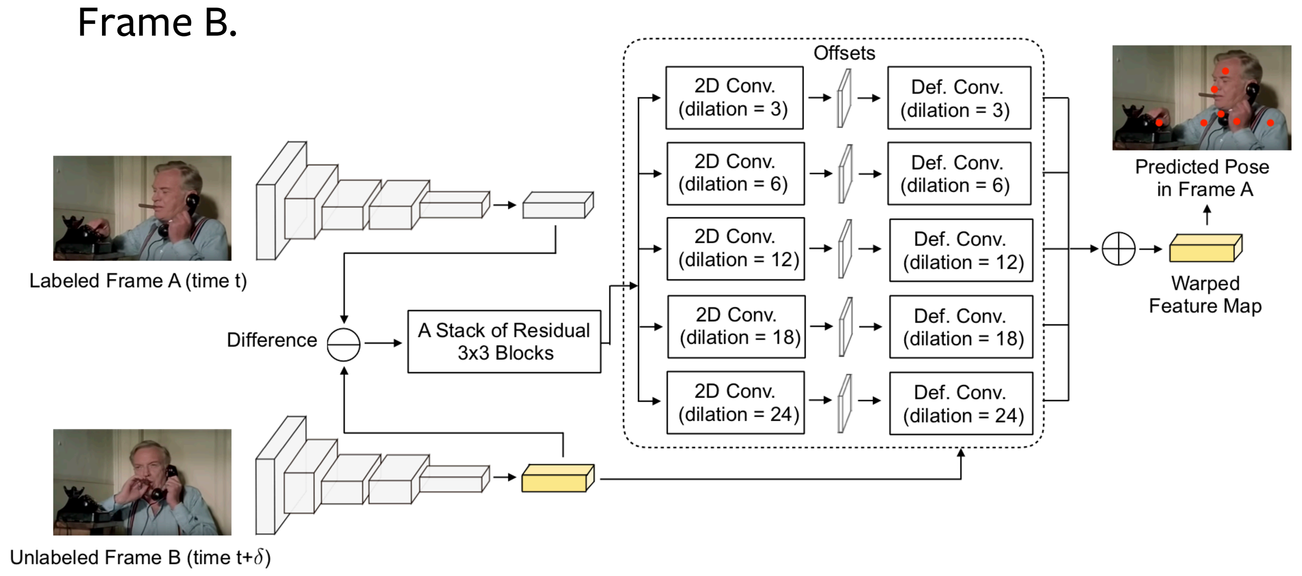
- We introduce the PoseWarper network that operates on sparsely annotated videos, i.e., pose annotations are given only every *k* frames.
- Given a pair of frames from the same video—a labeled Frame A and an unlabeled Frame B—we train our model to detect pose in Frame A using the features from Frame B.



II. The PoseWarper Network

Architecture

- We leverage dilated deformable convolutions to learn how to warp the pose heatmaps from an unlabeled Frame B to a labeled frame A.
- The warped heatmaps from Frame B are then used to detect Pose in a labeled Frame A.
- Our learned offsets implicitly learn motion cues between Frame A and Frame B.



III. Applications of the PoseWarper Network

Video Pose Propagation

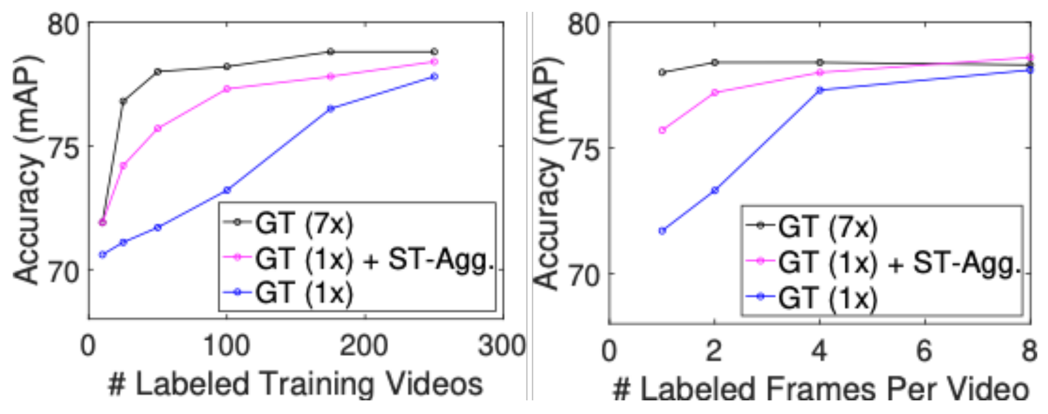
- Our goal is to propagate ground truth pose annotations across the entire video from only a few labeled frames.
- During inference, we can reverse the application direction of our trained model, i.e. warp ground truth pose from a labeled frame A to an unlabeled Frame B.

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

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Pseudo-labeling w/ HRNet [27]	79.1	86.5	81.4	74.7	81.4	79.4	72.3	79.3
Optical Flow Propagation (Farneback [55])	76.5	82.3	74.3	69.2	80.8	74.8	70.1	75.5
Optical Flow Propagation (FlowNet2 [29])	82.7	91.0	83.8	78.4	89.7	83.6	78.1	83.8
PoseWarper (no dilated convs)	86.1	91.7	88.0	83.5	90.2	87.3	84.6	87.2
PoseWarper (1 dilated conv)	85.0	91.6	88.0	83.7	89.6	87.3	84.7	87.0
PoseWarper (2 dilated convs)	85.8	92.4	88.8	84.9	91.0	88.4	86.0	88.0
PoseWarper (3 dilated convs)	86.1	92.6	89.2	85.5	91.3	88.8	86.3	88.4
PoseWarper (4 dilated convs)	**86.3**	92.6	**89.5**	85.9	**91.9**	88.8	86.4	88.6
PoseWarper (5 dilated convs)	86.0	**92.7**	**89.5**	**86.0**	91.5	**89.1**	**86.6**	**88.7**

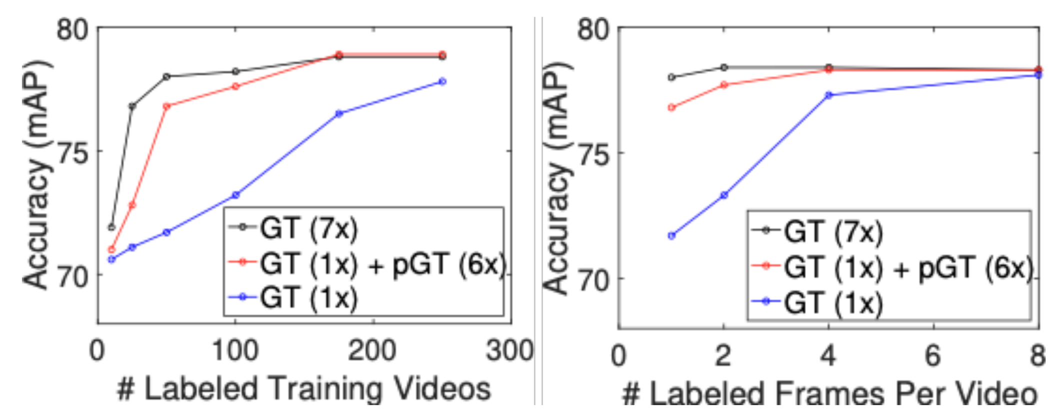
Spatiotemporal Pose Aggregation at Inference

- We can also use our learned warping mechanism to aggregate pose information from nearby frames during inference.
- This renders our approach more robust to occlusions, motion blur, video defocus, and rare poses.



Data Augmentation with PoseWarper

- We augment sparsely labeled video data with our propagated poses as pseudo ground truth labels.
- We then train a standard HRNet-W48 pose detector on this joint data.



IV. Additional Experiments

Comparison to State-of-the-Art

- We train our PoseWarper model on the full PoseTrack dataset, i.e., when frames in videos are densely labeled.
- During inference, we use our spatiotemporal pose aggregation scheme to aggregate information from 5 nearby frames.
- This allows us to achieve state-of-the-art pose detection results on PoseTrack17 and PoseTrack18 datasets.

Dataset	Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PoseTrack17 Val Set	Girdhar et al. [48]	72.8	75.6	65.3	54.3	63.5	60.9	51.8	64.1
	Xiu et al. [56]	66.7	73.3	68.3	61.1	67.5	67.0	61.3	66.5
	Bin et al [23]	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
	HRNet [27]	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
	MDPN [57]	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
	PoseWarper	81.4	88.3	83.9	78.0	82.4	80.5	73.6	**81.2**
PoseTrack17 Test Set	Girdhar et al. [48]	-	-	-	-	-	-	-	59.6
	Xiu et al. [56]	64.9	67.5	65.0	59.0	62.5	62.8	57.9	63.0
	Bin et al [23]	80.1	80.2	76.9	71.5	72.5	72.4	65.7	74.6
	HRNet [27]	80.1	80.2	76.9	72.0	73.4	72.5	67.0	74.9
	PoseWarper	79.5	84.3	80.1	75.8	77.6	76.8	70.8	**77.9**
PoseTrack18 Val Set	AlphaPose [58]	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
	MDPN [57]	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
	PoseWarper	79.9	86.3	82.4	77.5	79.8	78.8	73.2	**79.7**
PoseTrack18 Test Set	AlphaPose++ [57, 58]	-	-	-	66.2	-	-	65.0	67.6
	MDPN [57]	-	-	-	74.5	-	-	69.0	76.4
	PoseWarper	78.9	84.4	80.9	76.8	75.6	77.5	71.8	**78.0**

Interpreting our Learned Offsets

- Understanding what information is encoded in our learned offsets is challenging due to high dimensionality of the offsets, i.e., we are predicting 17x3x3=153 (x,y) displacements for every pixel.
- It appears that different offset maps encode different motion, thus performing a sort of motion decomposition of discriminative regions in the video.

Frame t | Frame t+5 | Channel 123 (x,y) | Channel 99 (x,y) | Offset Magnitudes

V. Conclusions

- Our approach reduces the need for densely labeled video data, while producing strong pose detection performance.
- Our state-of-the-art results on PoseTrack17 and PoseTrack18 datasets also show that our PoseWarper is useful even when training videos are densely labeled.
- The source code and our trained models are available at:

……………………………………………………………