Denoising and Selecting Pseudo-Heatmaps for Semi-Supervised Human Pose Estimation

Zhuoran Yu*†, Manchen Wang*‡, Yanbei Chen, Paolo Favaro, Davide Modolo
AWS AI Labs
zhuoran.yu@wisc.com, {manchenw,yanbec,pffavaro,dmodolo}@amazon.com

Abstract

We propose a new semi-supervised learning design for human pose estimation that revisits the popular dual-student framework and enhances it two ways. First, we introduce a denoising scheme to generate reliable pseudo-heatmaps as targets for learning from unlabeled data. This uses multi-view augmentations and a threshold-and-refine procedure to produce a pool of pseudo-heatmaps. Second, we select the learning targets from these pseudo-heatmaps guided by the estimated cross-student uncertainty. We evaluate our proposed method on multiple evaluation setups on the COCO benchmark. Our results show that our model outperforms previous state-of-the-art semi-supervised pose estimators, especially in extreme low-data regime. For example with only 0.5K labeled images our method is capable of surpassing the best competitor by 7.22 mAP (+25% absolute improvement). We also demonstrate that our model can learn effectively from unlabeled data in the wild to further boost its generalization and performance.

1. Introduction

Deep neural networks have achieved remarkable success in the past decade thanks to the availability of large-scale annotated datasets such as ImageNet [11] for image classification and MS COCO [24] for object detection and human pose estimation. However, obtaining a fully-labeled dataset at scale is extremely expensive in terms of both time and budget. In the case of human pose estimation, which is of interest in this work, such a dataset is even more challenging to build as it requires fine-grained and accurate annotations.

To mitigate the annotation cost in human pose estimation, we explore the use of semi-supervised learning (SSL) [8, 58, 59], a field of research that has received increasing attention in the scientific community. Recent SSL techniques make it possible to build high-performing models by training them simultaneously on small labeled and large unlabeled datasets for image classification [21, 23, 37, 39, 42, 51] and object detection [17, 17, 25, 26, 40, 57].

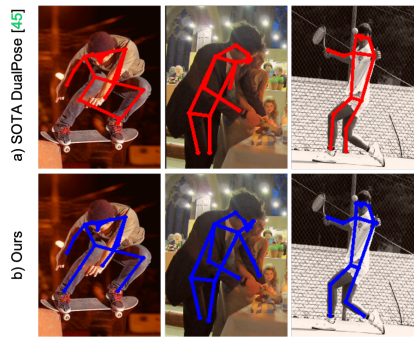

Figure 1. *Given person images in different challenging poses, an existing semi-supervised pose estimator tends to produce noisy targets on unlabeled images (top row). By contrast, we propose to generate learning targets of higher quality (bottom row) by ensembling the heatmaps from multiple augmented views and selecting more reliable targets guided by uncertainty.*

A key mechanism in recent SSL methods is the use of so-called *pseudo-labels*, i.e., the predictions of a model during its training, as a replacement for the ground truth targets. Much of the success of SSL methods is related to the quality of the pseudo-labels, as errors can lead to confirmation bias [1]. In the case of human pose estimation, pseudo-labels come in the form of 2D heatmaps (i.e., *pseudo-heatmaps*, Figure 2). Here, the use of SSL is made significantly more challenging by the difficulty of predicting heatmaps that are reliable across *all* keypoints, as shown in Figure 1.

To address these limitations we propose a novel training scheme for semi-supervised human pose estimation. Our method consists of several key designs, which significantly improve the quality of the heatmaps used as pseudo-labels, yielding state-of-the-art performance. In details, our method employs a dual-student framework [19] and enhances it with two major contributions: a denoising scheme to improve the quality of each student's pseudo-heatmaps

*These authors contributed equally to this work.
†Currently at The University of Wisconsin–Madison. Work conducted during an internship with AWS AI Labs.
‡Corresponding author.

and an uncertainty-guided pseudo-heatmap selection to determine the best possible pseudo-heatmaps to use as supervision to train the students.

Our *denoising* scheme improves the pseudo-heatmaps estimation in two ways: it ensembles the outputs of multiple strong and weak augmented views to obtain better estimates of each joint location and it refines the actual responses at these locations using a threshold-and-refine scheme. Furthermore, to deal with erroneous high confidence predictions from poorly calibrated models [33], we propose an uncertainty-based cross-student pseudo-heatmap *selection*. Inspired by recent works in semi-supervised classification [31, 37] and object detection [26, 53], our method discards noisy pseudo-labels by computing an approximate prediction uncertainty. Our solution differs from [26,37,53] in the way it parameterizes the uncertainty: instead of operating over simple one-hot vectors (classification) or bounding boxes coordinates (detection), we operate over full 2D heatmaps, one for each joint of a person's skeleton. To solve this complex problem, we propose to estimate the per-pixel uncertainty across all spatial locations. To the best of our knowledge we are the first to explore this direction for SSL human pose estimation.

To evaluate our design, we experiment using two versions of MS COCO [24]: *COCO-Partial* and *COCO-Additional*. Under both protocols (especially the former) our method achieves state-of-the-art performance and outperforms existing methods. For example, with 0.5K labeled instances, we outperform the strong DualPose [52] by 7.22 and 7.28 AP points (+25% absolute improvement) when evaluated with a single model and a model ensemble respectively, and consistently outperforms DualPose by 4-5 AP points with 1K and 2K labeled instances (+12%). With stronger input augmentation, we still consistently achieves more than 2 AP improvement over the best competitor. Finally, we conduct an extensive ablation study of the components of our design and validate their importance.

2. Related Work

Semi-supervised learning. Recent advances in semi-supervised learning have been achieved through various deep learning methods [9], with self-training [1,2,16,23,37, 39] and consistency regularization [3,4,21,38,42,50] being the most commonly used approaches. Self-training uses the model's own predictions to supervise the model itself. One well-known technique is pseudo-labelling [23], which converts model predictions to one-hot pseudo-labels and uses them to learn from unlabeled data. Consistency regularization [3, 4, 21, 38, 42, 50], on the other hand, enforces consistent predictions across input or model perturbations. State-of-the-art SSL methods usually combine these two prominent techniques [39, 56]. For example, FixMatch [39] uses weakly augmented views to generate pseudo-labels which are used as targets on the strongly augmented views to ensure consistently regularized output. Although this weak-strong data augmentation paradigm works well on image classification [39, 50] and object detection [25, 53, 57], weakly augmented views does not always give reliable learning targets for human pose estimation. This motivates us to propose more advanced formulations that generate more reliable pseudo-heatmaps for semi-supervised learning.

Semi-supervised human pose estimation. Unlike research in semi-supervised 3D human pose estimation [18, 29, 35], research in semi-supervised 2D human pose estimation just starts to attract more interest recently [19]. Due to the natural difference between 3D and 2D tasks, those 3D semi-supervised methods cannot be directly transferred to the 2D scenarios. Early work [44] on label-efficient 2D human pose estimation combines semi-supervised and weakly-supervised schemes. Several works develop semi-supervised methods for key-point localization [14, 30, 46]. For instance, Moskvyak et al [30] uses semantic consistency constraints to regularize the network whereas PLACL [46] combines curriculum learning with reinforcement learning to improve the training. Recently, DualPose [52] establishes a benchmark for semi-supervised human pose estimation and presents the dual student framework [5, 19] based on the weak-strong data augmentation paradigm [39, 50]. In this work, we focus on developing semi-supervised methods for 2D human pose estimation, but our method can also be adapted to other key-point localization tasks.

Uncertainty estimation in semi-supervised learning (SSL). Several research efforts have been made on the uncertainty estimation in network predictions [6,12,22,27,45]. In the SSL domain, [37] use Monte Carlo Dropout and measure uncertainty by calculating standard deviation of output probabilities, which helps to choose a better subset of pseudo-labels for classification. [55] use comparable techniques but employ predictive entropy in the context of medical segmentation tasks, while uncertainty is used as weights to compute pseudo-labels in [48]. In object detection, [53] samples jittered box to measure localization uncertainty as the box regression variance, meanwhile [26] explicitly model the box boundary uncertainty via an extra auxiliary branch. Our work differs from these existing works in two ways. First, we propose to estimate uncertainty using pixel-level Gaussian heatmap regression. Second, our estimator is computationally efficient – no new parameters are required for learning, which makes it scalable to estimate per-pixel uncertainty across all spatial locations.

3. Methodology

As Figure 2 shows, given an input image (either labeled or unlabeled), our semi-supervised pose estimator aug-

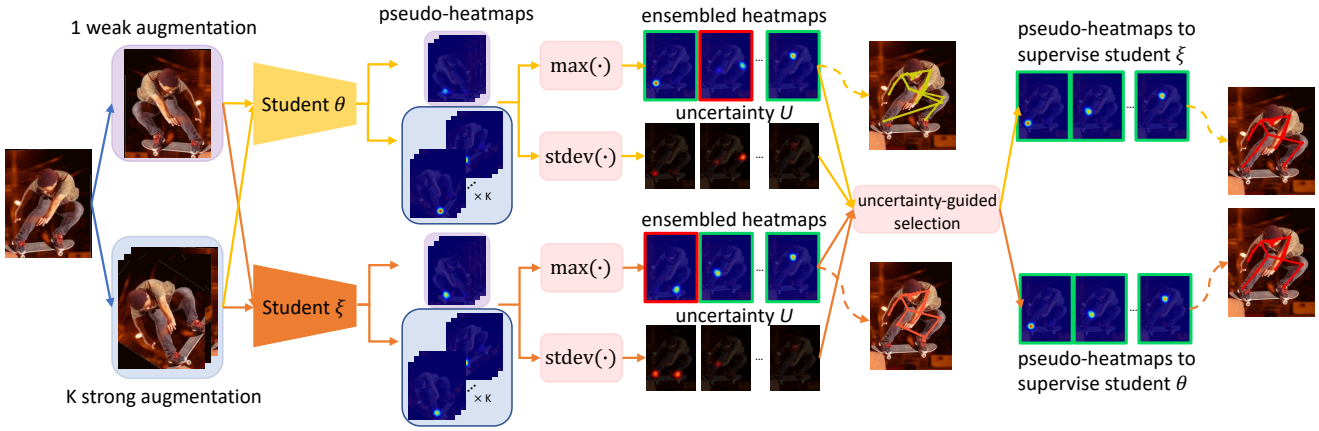

Figure 2. ***Our proposed semi-supervised pose estimation model.*** *Given an input image, we apply multi-view augmentation to generate a set of pseudo heatmaps, and ensemble these pseudo-heatmaps to obtain more precise pseudo groundtruth from each student network (Sec 3.2). To select the better pseudo groundtruth among two student networks as learning targets on unlabeled data, we introduce an uncertainty estimator to estimate the uncertainty of each pseudo heatmap and guide the selection of pseudo heatmaps (Sec 3.3).*

ments the image into different views, and generates pseudo-heatmaps based on multiple augmented views. The pseudo groundtruth of human pose is further produced guided by uncertainty. In the following, we first define our task and revisit a dual-student framework (Sec 3). We detail our approach which aggregates multiple augmented views to calibrate the pseudo-heatmaps (Sec 3.2), and select the more reliable learning targets guided by uncertainty (Sec 3.3).

3.1. Problem Definition and Preliminaries

Problem Definition. We consider the problem of semi-supervised human pose estimation. Given a small set of annotated images $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{M}$ (where x_i is the image labeled with pose annotations y_i) and a large set of unlabeled images $\mathcal{D}_u = \{u_i\}_{i=1}^{N}$ (where u_i is the unlabeled image), our goal is to develop a semi-supervised learning (SSL) framework that trains an human pose estimator with $\mathcal{D}_l$ and $\mathcal{D}_u$ jointly to achieve better model generalization.

A typical supervised pose estimator [32, 40, 43] (with model parameters θ) is trained to predict a set of heatmaps $H_i = \{h_{i,j}\}_{j=1}^{J}$ of J different human joints for an input image x_i by optimizing the following mean square errors:

$$\mathcal{L}_\theta^S = ||y_i - H_i||^2 \tag{1}$$

where $\mathcal{L}_\theta^S$ is the supervised loss for training network θ. The key success of semi-supervised learning is to enable learning from unlabeled data with unsupervised loss terms. We detail how to formulate these loss terms in the following.

Revisit the dual student framework. Following existing SSL methods which are often built upon consistency regularization [3,4,21,38,42,50], a recent work introduces a dual student framework [19, 52] for semi-supervised pose estimation, known as DualPose [52]. DualPose trains two student networks θ and ξ jointly, and employs the output from one network as the learning targets to supervise the other network. Specifically, an input image I is processed with weak-strong augmentation [39, 50] through affine transformations to obtain its weakly and strongly augmented counterparts I_w, I_s. Each student network predicts a set of heatmaps for the augmented inputs I_w, I_s:

$$\begin{aligned} H_{\theta,w} &= \theta(I_w) \quad \text{and} \quad H_{\theta,s} = \theta(I_s), \\ H_{\xi,w} &= \xi(I_w) \quad \text{and} \quad H_{\xi,s} = \xi(I_s), \end{aligned} \tag{2}$$

where $\theta(\cdot), \xi(\cdot)$ are two student networks. $H_{\theta,w}, H_{\xi,w} \in \mathbb{R}^{J\times h\times w}$ are predictions on the weakly augmented view I_w (where J is the number of joints, $h \times w$ denotes the spatial dimensions of a heatmap). $H_{\theta,s}, H_{\xi,s} \in \mathbb{R}^{J\times h\times w}$ are predictions on the strongly augmented view I_s.

To derive the learning targets on unlabeled images, the heatmaps predicted on weakly augmented views are used as the pseudo ground-truth (i.e., *pseudo-heatmaps*). The unsupervised loss terms are then computed as follows:

$$\begin{aligned} \mathcal{L}_\theta^U &= ||H_{\xi,w} - H_{\theta,s}||^2 = ||H_{\xi,w} - \theta(I_s)||^2 \\ \mathcal{L}_\xi^U &= ||H_{\theta,w} - H_{\xi,s}||^2 = ||H_{\theta,w} - \xi(I_s)||^2 \end{aligned} \tag{3}$$

where $\mathcal{L}_\theta^U, \mathcal{L}_\xi^U$ denote unsupervised loss terms for student networks θ, ξ. $H_{\xi,w}, H_{\theta,w}$ are used as the learning targets that supervise the networks to learn from unlabeled data.

Limitations. While being simple, this framework can produce unreliable learning targets on unlabeled data. Its pseudo-heatmaps can be noisy, as they are estimated from a single random weakly augmented view (Figure 1 left). As a consequence, the model may fit on this noise and exacerbate the training error propagation. In the next two sections we propose two novel formulations to address these problems.

3.2. Denoising pseudo-heatmaps

Using the outputs on weakly augmented views as learning targets for strongly augmented views was shown to be effective for semi-supervised learning in image classifica-

tion [25, 39, 50, 57]. However, in semi-supervised pose estimation, we find that the pseudo-heatmaps produced with one weakly augmented view are sometimes inaccurate (especially on some more challenging joints) and tend to have low responses (see Figure 1 top). To mitigate this problem, we propose two denoising solutions: first, we ensemble the outputs of multiple strong and weak augmented views to obtain better estimates of each joint location; and second, we refine the actual responses at these locations. We present the details of these solutions next.

Multi-view augmentation. To improve the pseudo-heatmaps estimation, we apply K different strong affine transformations on the input images to produce more variants for the same input image $\{I_{s,k}\}_{k=1}^{K}$. Our key insight is that by adding richer stochastic perturbations in the input space, we can cancel out mistakes and noises in learning targets caused by the randomness in one single weakly augmented view. This shares the same spirit as model ensembling [7, 28], where a committee of models is used to cover different regions of the version space and improve the accuracy of the final predictions. Given multiple strongly augmented views $\{I_{s,k}\}_{k=1}^{K}$ of the input I, we obtain a set of candidate pseudo-heatmaps $\{H_{\theta,s}^{k}\}_{k=1}^{K}$, which can be further ensembled to derive more reliable learning targets. Following a winner-take-all strategy, we compute the ensembled pseudo-heatmaps by taking the maximum scores per pixel over outputs of all augmented views, i.e., $\{H_{\theta,s}^{k}\}_{k=1}^{K}$ and $H_{\theta,w}$. That is, the ensembled pseudo-heatmaps from the two student networks are

$$\begin{aligned} P_\theta &= \max\{H_{\theta,w}, \{H_{\theta,s}^{k}\}_{k=1}^{K}\}, \\ P_\xi &= \max\{H_{\xi,w}, \{H_{\xi,s}^{k}\}_{k=1}^{K}\}, \end{aligned} \tag{4}$$

where P_θ, P_ξ are the pseudo-heatmaps obtained from networks θ, ξ, which are calibrated with better accuracy thanks to ensembling multiple outputs of $K+1$ augmented views.

A threshold-and-refine scheme. We employ a "threshold-and-refine" scheme to further denoise the pseudo-heatmaps P_θ, P_ξ in Eq. (4) in cleaner responses. This scheme is similar to the confidence-based thresholding scheme used to generate pseudo labels in existing semi-supervised image classification methods such as Pseudo-Label [23] and FixMatch [39], but we specially design it for generating pseudo-heatmaps in human pose estimation. Specifically, if the maximum response of pseudo-heatmaps P_θ is above a pre-defined threshold τ, we refine P_θ by applying a 2D Gaussian centered at the location corresponding to the maximum response of P_θ, where the 2D Gaussian is a special operation used in human pose estimation to translate ground truth points into 2D heatmaps.

With our threshold-and-refine scheme, we can derive refined pseudo-heatmaps P_θ, P_ξ to serve as more accurate unsupervised targets for training two student networks:

$$\begin{aligned} \mathcal{L}_\theta^U &= ||P_\xi - H_{\theta,s}||^2 = ||P_\xi - \theta(I_s)||^2, \\ \mathcal{L}_\xi^U &= ||P_\theta - H_{\xi,s}||^2 = ||P_\theta - \xi(I_s)||^2, \end{aligned} \tag{5}$$

where the above unsupervised loss terms constrain the output $\theta(I_s), \xi(I_s)$ of one strongly augmented view I_s. For both student networks, P_θ, P_ξ in Eq. (5) are more accurate pseudo-heatmaps compared to the heatmaps $H_{\theta,w}, H_{\xi,w}$ in Eq. (3) generated from one weakly augmented views.

3.3. Uncertainty-guided pseudo-heatmaps selection

The two student networks θ and ξ can perform differently, even when predicting the same joint in the same image. Importantly, one student may output more reliable learning targets than the other student and we can exploit this information to further improve the final pseudo-heatmaps used for cross-training. For this, we propose to select the most accurate pseudo-hatmaps using an uncertainty-guided selection scheme that rejects the heatmaps with higher uncertainty scores.

Uncertainty estimation on heatmaps. Given an unlabeled image I and a set of heatmaps $\{H_{\theta,w}, \{H_{\theta,s}^{k}\}_{k=1}^{K}\}$ predicted on its augmented views by student network θ, we can estimate the uncertainty of the heatmap for each single joint. Rather than following the prior work [37] that uses Monte Carlo Dropout [12] for uncertainty estimation, we propose a novel uncertainty estimator which computes the uncertainty based on the outputs of multiple augmented views. Specifically, the uncertainty is estimated as the pixel-wise standard deviation across the set of heatmaps predicted on multiple augmented views for each joint. The uncertainty can be expressed as follows for student networks θ, ξ:

$$\begin{aligned} U_\theta &= \text{stdev}(\{H_{\theta,w}, \{H_{\theta,s}^{k}\}_{k=1}^{K}\}), \\ U_\xi &= \text{stdev}(\{H_{\xi,w}, \{H_{\xi,s}^{k}\}_{k=1}^{K}\}), \end{aligned} \tag{6}$$

where $U_\theta, U_\xi \in \mathbb{R}^{J\times h\times w}$ have the same dimensions as $H_{\theta,w}$, $H_{\theta,s}^{k}$, $H_{\xi,w}$, $H_{\xi,s}^{k}$, e.g., $U_\theta{=}\{U_j\}_{j=1}^{J}$ is a set of uncertainty maps for J different joints. To further derive a scalar value that represents the uncertainty of each heatmap, we take the maximum value on uncertainty maps, which leads to a set of J uncertainty scores: $u_\theta{=}\{u_j\}_{j=1}^{J}{=}\{\max(U_j)\}_{j=1}^{J}$, where $\max(U_j)$ is the maximum uncertainty value of the j_{th} joint.

Uncertainty-guided selection across student networks. Given the two set of uncertainty maps U_θ, U_ξ which measure the reliability of the two set of heatmaps produced by two student networks, we can select the more reliable outputs as pseudo-heatmaps for the unlabeled images. Specifically, we trust one student's own output over the other student when its uncertainty scores are lower than a certain margin Δ, than those of the other student. The learning tar-

gets selected by uncertainty for student network θ are:

$$\hat{P}_{\theta,j} = \begin{cases} P_{\theta,j}, & \text{if } u_{\theta,j} + \Delta < u_{\xi,j} \\ \mathcal{P}_{\xi,j}, & \text{otherwise} \end{cases} \quad (7)$$

where $j \in [1, J]$ denotes the index of J different joints. $\hat{P}_{\theta,j}$ is the final pseudo-heatmap of the j_{th} joint which is selected between $P_{\theta,j}$ and $P_{\xi,j}$ based on their uncertainty score $u_{\theta,j}$ and $u_{\xi,j}$. If the uncertainty $u_{\theta,j}$ of $P_{\theta,j}$ is lower than the uncertainty $u_{\xi,j}$ of $P_{\xi,j}$ than a certain margin Δ, then $P_{\theta,j}$ is selected as pseudo-heatmap for the j_{th} joint; otherwise $P_{\xi,j}$ is selected. We compute the learning targets $\hat{P}_{\xi,j}$ for student network ξ in an equivalent way as Eq. (7). Finally, from these we can obtain more reliable pseudo-heatmaps $\hat{P}_\theta, \hat{P}_\xi \in \mathbb{R}^{J \times h \times w}$ for J different joints to serve as the final pseudo-groundtruth for the student networks θ and ξ.

Semi-supervised learning objective. With P_θ, P_ξ as the learning targets, we can now formulate the unsupervised loss on unlabeled data on two student networks θ, ξ, in the similar spirit as Eq. (5):

$$\mathcal{L}_\theta^U = ||H_{\theta,s} - \hat{P}_\xi||^2, \quad \mathcal{L}_\xi^U = ||H_{\xi,s} - \hat{P}_\theta||^2, \quad (8)$$

where $\hat{P}_\theta, \hat{P}_\xi$ are the targets selected guided by uncertainty as in Eq. (7). The final semi-supervised learning objective is to jointly optimize the above unsupervised loss along with the supervised loss in Eq. (1):

$$\mathcal{L}_\theta = \mathcal{L}_\theta^S + \mathcal{L}_\theta^U, \quad \mathcal{L}_\xi = \mathcal{L}_\xi^S + \mathcal{L}_\xi^U, \quad (9)$$

where $\mathcal{L}_\theta^S, \mathcal{L}_\xi^S$ are the supervised loss computed on the labeled images based on Eq. (1). $\mathcal{L}_\theta^U, \mathcal{L}_\xi^U$ are the unsupervised loss computed on unlabeled images based on Eq. (8).

4. Experiments

4.1. Datasets and Evaluation

COCO Dataset. COCO is a large-scale benchmark dataset for object detection, segmentation and human pose estimation. For semi-supervised human pose estimation, we use the following sets. **train2017** contains 118K images and 150K labeled person images with 17 keypoints; **unlabeled2017** contains 123K unlabeled images; **val2017** contains 5K images and 6K labeled person images; **test-dev** set contains 20K images where annotations are private (we obtain results by submitting our predictions to their evaluation server). Unless explicitly stated, we evaluate the models on the val2017. We validate the performance of our method under two evaluation protocols: *COCO-Partial* and *COCO-Additional*. *COCO-Partial* splits train2017 into 0.5K, 1K, 2K, 5K, 10K incremental labeled sets and uses the remaining images as unlabeled. *COCO-Additional* uses the whole train2017 as labeled data and unlabeled2017 as unlabeled.

AI Challenger Dataset. To demonstrate the effectivness of our method in leveraging additional unlabeled data, we also experiment with the large-scale AI Challenger Dataset [47], consisting of 210k images. While these images come with annotations, we disregard them as consider them unlabelled.

Evaluation metrics. We report mean average precision (AP) over 10 Object Keypoint Similarity (OKS) [24]'s thresholds: [0.5, 0.55 ..., 0.9, 0.95]. OKS is calculated as the Euclidean distances between each corresponding ground truth and the detected keypoints, normalized by the scale of the person. To measures each keypoint's localization accuracy, we also report PCK (Percentage of Correct Keypoints) [54] score. A joint is correct is it falls within αl pixels of the ground-truth position, where α is a constant and l is maximum side length of the ground-truth person bounding box. The PCK@0.1 (α = 0.1) score is reported.

4.2. Implementation details

Benchmark settings. For fair comparison, we implement all models (baseline and ours) in the same codebase and compare them using the same backbones. On *COCO-Partial*, we evaluate each method with 3 random seeds and report the mean and standard deviation across the 3 runs. We follow [52] and use Simple Baseline [49] or HRNet [41] as the pose estimator and conduct our experiments with various backbones. Specifically, for *COCO-Partial*, we use ResNet18 [13] for the low-data regimes (0.5K, 1K and 2K) and use ResNet50 for high-data regimes (5K and 10K). Since performance on ResNet18 gets saturated with more labeled data, we also use ResNet50, ResNet101, ResNet152 and HRNetW48 for evaluation.

Training details. Our backbone network is initialized by pre-training on ImageNet [11]. The base learning rate is 0.001 and Adam [20] optimizer is used. For *COCO-Partial*, we train our model for 100 epochs and decay learning rate by a factor of 10 at the 70-th and 90-th epoch. For *COCO-Additional* we train our model for 400 epochs and decay learning rate at the 300-th and 350-th epoch. We set $\Delta = 0.05$ as the margin between uncertainties of two students.

Data Augmentation. Our *weak* augmentation policy[1] uses random affine (A) transformations with rotation degrees sampled from $[-30°, 30°]$ and scale factors sampled from $[0.75, 1.25]$. Our *strong* augmentation policy samples rotation degrees from $[-60°, 60°]$ and scale factors from $[0.5, 1.5]$ with Beta distribution ($\alpha = \beta = 0.75$). Finally, when stated, we also train with JointCutout (JC) [52].

Testing details. We apply a two-stage top-down paradigm similar to [10, 34]. On COCO val2017, we run the pose estimators on ground truth bounding boxes without image flipping. On COCO test-dev set, we use the popular person detector of Simple Baseline [49] (AP of 60.9). We follow previous works [10, 32] and predict joint locations as the

[1] We keep our weak augmentations the same as DualPose [52] to ensure that our comparison is fair and can highlight the contribution of our model in selecting more reliable pseudo heatmaps for semi-supervised learning.

Method	Aug	0.5K	1K	2K	5K	10K
		ResNet18			ResNet50	
Supervised Baseline	A	22.05 ± 1.12	30.91 ± 0.64	36.07 ± 0.50	49.31 ± 0.44	55.98 ± 0.08
DataDistill [36]	A	-	37.6	-	51.6	56.6
DualPose [52]	A	32.16 ± 1.18	41.54 ± 0.66	46.48 ± 0.44	58.39 ± 0.53	63.07 ± 0.47
Ours	A	**39.38 ± 0.94** (+7.22)	**46.27 ± 0.50** (+4.73)	**50.74 ± 0.28** (+4.26)	**60.67 ± 0.12** (+2.28)	**63.81 ± 0.32** (+0.74)
DualPose-Ensemble [52]	A	32.98 ± 1.27	42.67 ± 0.67	48.19 ± 0.45	59.37 ± 0.44	64.46 ± 0.46
Ours-Ensemble	A	**40.26 ± 0.87** (+7.28)	**47.32 ± 0.70** (+4.65)	**51.96 ± 0.35** (+3.77)	**61.62 ± 0.17** (+2.25)	**64.84 ± 0.25** (+0.38)
SSPCM [15]	A+CO	-	46.9	-	61.6	**65.4**
DualPose [52]	A+JC	36.89 ± 0.20	44.97 ± 0.14	48.67 ± 0.16	60.62 ± 0.25	64.25 ± 0.37
Ours	A+JC	**42.13 ± 0.21** (+5.24)	**47.58 ± 0.29** (+2.61)	**51.25 ± 0.27** (+2.58)	**62.14 ± 0.10** (+1.52)	**65.36 ± 0.15** (+1.11)
DualPose Ensemble [52]	A+JC	37.56 ± 0.47	46.19 ± 0.14	50.48 ± 0.26	62.04 ± 0.17	65.69 ± 0.21
Ours-Ensemble	A+JC	**43.06 ± 0.20** (+5.50)	**48.70 ± 0.09** (+2.51)	**52.54 ± 0.32** (+2.06)	**63.34 ± 0.02** (+1.30)	**66.46 ± 0.15** (+0.77)

Table 1. ***Comparison on COCO-Partial***. *We report results using different number of labeled images (i.e., 0.5K, 1K, 2K, 5K, 10K), and test the models under different augmentation strategies: 'A' means affine transformation; 'JC' denotes JointCutout [52]; 'CO' denotes Cut-Occlude [15]. 'Ensemble' means averaging two student networks' output. Metric: AP. The best result in each setup is in* **bold**. *We show the improved margins over the best competitor DualPose [52] in red.*

average between the original and flipped images.

4.3. Comparison with the state-of-the-art

Comparison on *COCO-Partial protocol*. We compare four semi-supervised pose estimation models: DataDistill [36], DualPose [52], SSPCM [15] and ours on COCO val2017, in terms of OKS-based AP. Table 1 demonstrates the stronger performance of our model and its ability to leverage more reliable pseudo-heatmaps to achieve better generalization: our model consistently outperforms the best competitor DualPose, under all setups, including using different number of labeled images (ranging from 0.5K to 10K), different backbone networks (ResNet18 and ResNet50), and different types of data augmentation strategies (A and A+JC). In particular, we find it performs especially well in the low-label regime. For example, when using 0.5K or 1K labeled images, our improved margins over DualPose are significantly larger, up to +7.22 when using a simple affine transformation. Considering the high complexity and prohibitive cost of annotating human pose estimation datasets, we believe this low-regime to be the most interesting target for SSL, as it would enable training budget-friendly real-world pose estimation models on only few hundred samples, while still achieving very competitive performance.

Evaluating individual joint predictions' quality. OKS-based AP computes aggregated statistics over multiple joints and a person's skeleton is considered correct if *most* of the predicted joints falls within a certain distance from their corresponding ground truth (i.e., a single joint mistake is irrelevant for OKS when the majority are correct). To truly understand the improvement that our model brings over DualPose, we now investigate the per-joint performance using the Percentage of Correct Keypoint (PCK) metric. As shown in Table 2, we achieve substantially better joint estimation performance across all 7 (meta-)joints (we average symmetric joints - left and right). Interestingly, the largest improvements can be observed on the challenging and dynamic limb joints (e.g., +8.1 on Wrist), showing the importance of modelling uncertainty.

Method	Ankl	Knee	Hip	Wrist	Elb	Shld	Head
DualPose [52]	61.4	62.8	62.5	62.1	68.9	77.2	91.7
Ours	**65.7**	**67.6**	**68.6**	**70.2**	**75.8**	**81.7**	**93.5**

Table 2. ***Comparison of per joint results on COCO-Partial.*** *Metric: PCK, which measures per-joint localization accuracy.*

Comparison on *COCO-Additional protocol*. To study the effect of training on datasets of larger scale, *COCO-Additional* uses COCO train2017 as labeled set and COCO unlabeled2017 as unlabeled set, leading to a total of 118K labeled and 123k unlabeled images. To exploit unlabeled data in the wild, we also experiment by expanding the aforementioned unlabeled set with the AI Challenger (AIC) Dataset [47], resulting in a total of 333K unlabeled images. We experiment with different backbone networks (ResNet50, ResNet152) and compare with state-of-the-art competitors in Table 3. We find that training with unlabeled data from COCO-unlabeled2017 and AI Challenger gives the best results for all backbones. For instance, when using ResNet50 as backbone, our model achieves 73.3 AP which is better than 72.3 by DualPose.

4.4. Ablation Study

In this section, we now ablate the components of our model. Unless specified, we use the *COCO-Partial* protocol with 1K labeled images and evaluate on COCO val2017. We first provide model ablation in Section 4.4.1, and then analyze individual components in Section 4.4.2 and 4.4.3.

4.4.1 Model ablation study

Different model components. In Table 4, we evaluate different components of our approach and ablate how each of

Method	Unlabeled Data	Backbone	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR
SimpleBaseline [49]	-	ResNet50	70.2	90.9	78.3	67.1	75.9	75.8
SB w/ DualPose [52]	COCO-unlabeled2017		72.3	91.8	80.5	69.3	77.8	77.7
SB w/ Ours	COCO-unlabeled2017		72.5	91.8	81.0	69.7	77.9	77.8
SB w/ Ours	COCO-unlabeled2017 + AIC		**73.3**	**92.1**	**82.0**	**70.9**	**78.5**	**78.9**
SimpleBaseline [49]	-	ResNet152	71.9	91.4	80.1	68.9	77.4	77.5
SB w/ DualPose [52]	COCO-unlabeled2017		73.7	92.1	82.1	71.0	79.0	79.1
SB w/ Ours	COCO-unlabeled2017		73.8	91.9	82.1	71.1	79.2	79.2
SB w/ Ours	COCO-unlabeled2017 + AIC		**74.2**	**92.1**	**82.4**	**71.5**	**79.6**	**79.4**

Table 3. ***Comparison on COCO-Additional with the test-dev set.*** *The entire COCO training set (train2017) is used as labeled sets. COCO-unlabeled2017 set (and AI Challenger (AIC) [47]) is used as unlabeled set. The person detection results are provided by Simple Baseline (SB) [49] and flipping strategy is used.*

	DualPose	Ours		
Multi-View Augmentation		✓	✓	✓
Threshold-and-Refine			✓	✓
Uncertainty Guided Selection				✓
AP	42.67	44.91	46.49	**47.97**

Table 4. ***Ablation study of different model components.***

them contributes to the model performance. Building upon DualPose, we find that multi-view augmentation improves the performance by 2.24 points. Denoising with threshold-and-refine brings additional 1.58 AP. Further applying the uncertainty-guided pseudo-heatmaps selection, the performance reaches 47.97 AP, which is 5.3 points better than DualPose. This suggests the collective effects of different model components.

Different backbones and more unlabeled data. In Table 5, we train our model using varying amounts of unlabeled data (i.e., COCO unlabeled and COCO unlabeled + AIC) and compare them with a fully-supervised model solely trained on COCO train2017. As Table 5 shows, we find that our approach outperforms the supervised baseline using all backbones, achieving a noteworthy increase of up to +3.2 AP when only using COCO unlabeled as unlabeled set. By using the additional AIC unlabeled dataset, our method further improves the performance by an additional margin, up to +1.4 AP. This shows that our model is capable of taking advantage of additional unlabelled data and further improve the performance of human pose estimation.

4.4.2 Analysis of denoising pseudo-heatmaps

Effect of multi-view augmentation. As described in Sec 3.2, we augment each image into K strong and 1 weak views. In Table 6 we evaluate this choice against baselines using only 1, K and $K+1$ weak views. We also compare two ways of aggregating the $K+1$ pseudo-heatmaps obtained from augmented views: average and maximum response. Results show that taking the maximum response is better than average. Furthermore, adding K additional augmentation (either weak or strong) always greatly improves the model performance. This shows the benefits of introducing multi-view augmentation, as ensembling outputs from multiple views tends to cancel out the individual errors in single view, providing more accurate pseudo-heatmaps. Lastly, we find that using K strong views is more effective than using K weak ones, which improves AP from 43.74 to 44.91 (+1.17). This is because strong augmentations provide more diverse variations in input space, leading to more precise ensembled pseudo-heatmaps.

Method	Backbone	Unlabeled Data	AP	AP_{75}
Supervised	ResNet50	-	70.9	78.2
Ours		COCO-unlabeled2017	74.1	81.5
Ours		COCO-unlabeled2017+AIC	**75.5**	**82.7**
Supervised	ResNet101	-	72.5	80.3
Ours		COCO-unlabeled2017	75.7	83.6
Ours		COCO-unlabeled2017+AIC	**76.6**	**84.6**
Supervised	ResNet152	-	73.2	81.2
Ours		COCO-unlabeled2017	76.0	83.7
Ours		COCO-unlabeled2017+AIC	**76.7**	**84.6**
Supervised	HRNetW48	-	77.2	84.6
Ours		COCO-unlabeled2017	79.4	86.7
Ours		COCO-unlabeled2017+AIC	**79.8**	**86.9**

Table 5. ***Ablation study of additional unlabeled data.*** *Using more unlabeled data consistently improves model performance.*

Row	Augmentation	Aggregate	mAP
1	1 weak	N/A	42.67
2	(1 + K) weak	`avg`	42.11
3	(1 + K) weak	`max`	43.74
4	K weak + 1 strong	`avg`	41.77
5	K weak + 1 strong	`max`	43.92
6	1 weak + K strong	`avg`	39.58
7	1 weak + K strong	`max`	**44.91**

Table 6. ***Effect of the multi-view augmentation.*** *Row 1 is the baseline with 1 weak augmentation. Row 2-3/4-5/6-7 compare models with K additional weak/strong augmentation. Note: we can aggregate multi-view pseudo heatmaps by taking the maximum response (i.e., `max`, Eq. (4)) or average response (i.e., `avg`).*

Analysis of strong augmentation over different joints. In Eq. (4) we obtain the maximum response heatmap across K strong and 1 weak augmented view. To understand the importance of strong views, we compute the percentage of pseudo-heatmaps selected from them for each joint independently (Figure 4). Interestingly, there is no unanimous consensus and the percentage varies from 30 (Head) to 60 (Ankle), which explains the results of Table 2: our method achieves similar performance as DualPose on Head (+1.8

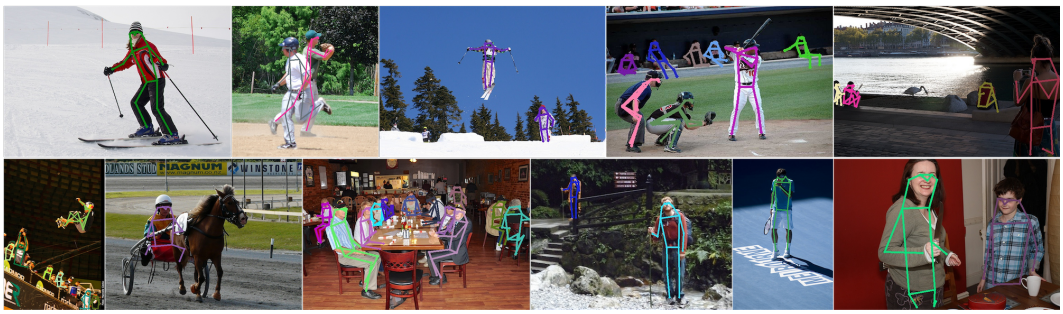

Figure 3. ***Qualitative results*** *showing example images in the COCO datasets, which covers persons in different viewpoints, appearance changes, and performing different activities. These results show that our model estimates human pose of good quality on unlabeled data.*

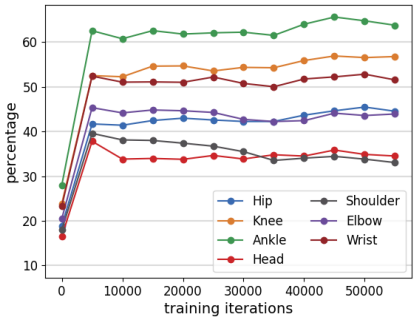

Figure 4. ***Analysis of pseudo-heatmaps selected from strongly augmented views***. *The model tends to select pseudo-heatmaps from weakly-augmented views for torso joints, while selecting pseudo-heatmaps from strongly-augmented views for limb joints.*

Row	Threshold	w/ Refinement	w/o Refinement
1	0	44.32	43.91
2	0.1	**47.97**	44.73
3	0.2	45.80	45.07
4	0.3	45.18	44.65

Table 7. ***Effect of the "threshold-and-refine" scheme.*** *From row 1 to 4, we increase the threshold when applying thresholding on pseudo heatmaps, and compare results with and without refining pseudo heatmaps with 2D Gaussion. Metric: AP.*

PCK), as they both rely on weak augmentations to learn, but it improves error-prone limb joints by up to +8.1 PCK, thanks to its leveraging of strong augmentations.

Effect of "threshold-and-refine" scheme. We evaluate this scheme in Table 7. Our model achieves the lowest AP (43.91) when using neither thresholding nor refining. Moreover, refining pseudo heatmaps with 2D Gaussian always improves the model performance, especially with a threshold of 0.1 (AP 47.97, +4.06 over not using either). This shows the benefits of applying thresholding to remove the low response in pseudo heatmaps, and refining the heatmaps using 2D Gaussian to produce cleaner pseudo heatmaps.

4.4.3 Analysis of uncertainty-guided selection

As mentioned in Sec 3.3, we introduce an uncertainty estimator (Eq. (6)) to estimate the uncertainty per joint based on a set of heatmaps from multiple augmented views. We

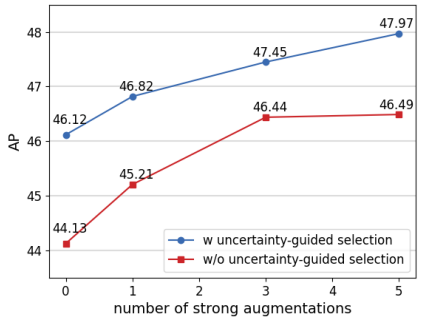

Figure 5. ***Effect of uncertainty-guided pseudo heatmaps selection.*** *We study our model with and without the uncertainty-guided selection when using different number of augmented views.*

also employ the uncertainty as an indicator to select more reliable pseudo groundtruth among two student networks to learn from unlabeled data (Eq. (7)). To study their effect, we conduct an ablation study by removing the uncertainty-guided selection component from our model (Figure 5). We observe that using the uncertainty estimator to select the pseudo-heatmaps (blue curve) gives significant better performance than not using it (red curve). Interestingly, both curves (with and without uncertainty) improve their performance as we increase the number of augmented views, validating the previous hypothesis that multiple views are important and showcasing the complementarity of our multi-view and uncertainty contributions.

5. Conclusion

We presented a new semi-supervised learning approach for human pose estimation. We introduceed multi-view augmentation to generate a candidate pool of pseudo heatmaps and selected the more reliable pseudo heatmaps guided by uncertainty. We exploited the pseudo-heatmaps as groundtruth on unlabeled data to perform semi-supervised learning. Our experimental results show that our model outperforms the state-of-the-art human pose estimator, especially in the extreme low-label regime where we have only a small fraction of labeled data (e.g., 0.5K, 1K). We also show that our model can effectively exploit unlabeled data in the wild to further boost its performance.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, 2020. 1, 2

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *arXiv preprint arXiv:1412.4864*, 2014. 2

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *ICLR*, 2020. 2, 3

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2, 3

[5] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *CLT*, 1998. 2

[6] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, 2015. 2

[7] Leo Breiman. Random forests. *Machine learning*, 2001. 4

[8] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE TNN*, 2009. 1

[9] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE TPAMI*, 2022. 2

[10] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 5

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 5

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2, 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[14] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, 2018. 2

[15] Linzhi Huang, Yulong Li, Hongbo Tian, Yue Yang, Xiangang Li, Weihong Deng, and Jieping Ye. Semi-supervised 2d human pose estimation driven by position inconsistency pseudo label correction module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 693–703, 2023. 6

[16] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *CVPR*, 2019. 2

[17] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 1

[18] Atul Kanaujia, Cristian Sminchisescu, and Dimitris Metaxas. Semi-supervised hierarchical models for 3d human pose reconstruction. In *CVPR*, 2007. 2

[19] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *ICCV*, 2019. 1, 2, 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 2, 3

[22] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30, 2017. 2

[23] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, 2013. 1, 2, 4

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 5

[25] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1, 2, 4

[26] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *CVPR*, 2022. 1, 2

[27] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *NeurIPS*, 2018. 2

[28] Tom M Mitchell. Generalization as search. *Artificial intelligence*, 1982. 4

[29] Rahul Mitra, Nitesh B Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *CVPR*, 2020. 2

[30] Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. *arXiv preprint arXiv:2101.07988*, 2021. 2

[31] Subhabrata Mukherjee and Ahmed Awadallah. Uncertainty-aware self-training for few-shot text classification. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21199–21212. Curran Associates, Inc., 2020. 2

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3, 5

[33] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015. 2

[34] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 5

[35] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with tem-

poral convolutions and semi-supervised training. In *CVPR*, 2019. 2

[36] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *CVPR*, 2018. 6

[37] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. 1, 2, 4

[38] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *arXiv preprint arXiv:1606.04586*, 2016. 2, 3

[39] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 1, 2, 3, 4

[40] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1, 3

[41] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 5

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1, 2, 3

[43] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NeurIPS*, 2014. 3

[44] Norimichi Ukita and Yusuke Uematsu. Semi-and weakly-supervised human pose estimation. *CVIU*, 2018. 2

[45] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*. PMLR, 2020. 2

[46] Can Wang, Sheng Jin, Yingda Guan, Wentao Liu, Chen Qian, Ping Luo, and Wanli Ouyang. Pseudo-labeled auto-curriculum learning for semi-supervised keypoint localization. *arXiv preprint arXiv:2201.08613*, 2022. 2

[47] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Large-scale datasets for going deeper in image understanding. In *ICME*, 2019. 5, 6, 7

[48] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. 3d semi-supervised learning with uncertainty-aware multi-view co-training. In *WACV*, 2020. 2

[49] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 5, 7

[50] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, 2020. 2, 3, 4

[51] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 1

[52] Rongchang Xie, Chunyu Wang, Wenjun Zeng, and Yizhou Wang. An empirical study of the collapsing problem in semi-supervised 2d human pose estimation. In *ICCV*, 2021. 2, 3, 5, 6, 7

[53] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, 2021. 2

[54] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE TPAMI*, 2013. 5

[55] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, 2019. 2

[56] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *NeurIPS*, 2021. 2

[57] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *CVPR*, 2021. 1, 2, 4

[58] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 2009. 1

[59] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005. 1