An Empirical Study of the Collapsing Problem in Semi-Supervised 2D Human Pose Estimation

Rongchang Xie¹, Chunyu Wang², Wenjun Zeng², Yizhou Wang³

¹Center for Data Science, Peking University  ²Microsoft Research Asia  ³Center on Frontiers of Computing Studies, CS Dept., Peking University

{rongchangxie, yizhou.wang}@pku.edu.cn, {chnuwa, wezeng}@microsoft.com

Abstract

Most semi-supervised learning models are consistency-based, which leverage unlabeled images by maximizing the similarity between different augmentations of an image. But when we apply them to human pose estimation that has extremely imbalanced class distribution, they often collapse and predict every pixel in unlabeled images as background. We find this is because the decision boundary passes the high-density areas of the minor class so more and more pixels are gradually mis-classified as background. In this work, we present a surprisingly simple approach to drive the model to learn in the correct direction. For each image, it composes a pair of easy-hard augmentations and uses the more accurate predictions on the easy image to teach the network to learn pose information of the hard one. The accuracy superiority of teaching signals allows the network to be “monotonically” improved which effectively avoids collapsing. We apply our method to the state-of-the-art pose estimators and it further improves their performance on three public datasets. The source code and pretrained models have been released at https://github.com/xierc/Semi_Human_Pose.

1. Introduction

2D human pose estimation has many practical applications such as 3D pose modeling [48, 35, 23] and action recognition [37, 38]. The early works in deep learning try to regress joint coordinates from images directly [34, 5]. But most recent ones adopt the heatmap-based framework [33, 39, 24, 30, 41] because it provides better supervision. But there is a more important but less explored problem of learning robust models that perform well on unseen wild images. One solution is to fit the “whole” world by infinitely increasing training images. The other is to transfer pre-trained models to new domains by unsupervised fine-tuning. The common basis behind the two approaches is Semi-Supervised Learning (SSL) — how to leverage unlabeled images to obtain a generalizable model?

The previous SSL works have primarily focused on the classification task. In general, there are two strategies to explore unlabeled images. The first is Pseudo labeling [26, 42] which first learns an initial model on only labeled images in a supervised way. Then, for each unlabeled image, it applies the initial model to obtain hard or soft pseudo labels representing its category. Finally, it learns the ultimate model on the combined dataset of labeled and pseudo-labeled images. However, the performance of the method is largely limited by that of the initial model which is learned only on the labeled images and fixed thereafter.

The second class of methods [2, 19, 27, 28, 31] learn about unlabeled images by requiring the network to have similar predictions for different augmentations of the same image. They are better than the pseudo labeling methods because the accuracy is not limited by the fixed labeling network. However, when we apply them to 2D pose estimation, we find that all of them encounter the collapsing problem meaning that, within few training iterations, the models begin to predict every pixel in unlabeled images as
background. As a result, the prediction accuracy becomes even worse than the initial supervised model.

The collapsing problem is not identified as a serious issue in previous works because most of them were only evaluated on the well-balanced classification task. But we find it is vital for tasks with severe class imbalance such as human pose estimation, which has not received sufficient attention. It occurs because when the network makes different predictions on the corresponding pixels, it lacks sufficient information to determine the correct optimization path. Blindly minimizing their discrepancy causes the decision boundary to be incorrectly formed due to imbalance and pass through the high-density area of the minor class as revealed in [14]. It leads to the situation where a growing number of pixels are mis-classified as background.

In this work, a simple approach is presented to address the collapsing problem. We first introduce the concept of easy-hard augmentation pair and, by definition, a network should obtain better average accuracy on a certain dataset with easy augmentation than on the same dataset with hard augmentation. Then, for each unlabeled image, we compose an easy and a hard augmentation, feed them to the network and obtain two heatmap predictions. We use the accurate predictions on the easy augmentation to teach the network to learn about the corresponding hard augmentation (see Figure 1). However, the hard augmentation will not be used for teaching the network to learn about the easy augmentation, which avoids high response samples being pulled to background as illustrated in Figure 2. The relative accuracy superiority of the teaching signals allows the network to be “monotonically” improved which stabilizes the training and avoids collapsing.

Our approach is general and applies to most consistency-based SSL methods such as [17, 19] for stopping collapsing. We empirically validate it on a simple baseline as well as on the state-of-the-art method [17] which jointly learns two models. Both methods collapse in their original setting and our easy-hard augmentation strategy helps avoid the problem. We extensively evaluate them on three public datasets of COCO [22], MPII [1] and H36M [15]. When the number of labeled images is small, our approach increases the mean Average Precision (AP) by about 13% (from 31.5% to 44.6%) compared to the supervised counterpart which only uses labeled data for training. As a comparison, the pseudo labeling methods of [42] and [26] only get 37.2% and 37.6% mean AP, respectively. More importantly, when we apply our method to the best 2D pose estimator and use all available labeled training images, it can further improve the performance by a decent margin by exploring unlabeled images. We also report results when our approach is used for semi-supervised pre-training and domain adaptation tasks. The versatile practical applications in various settings validate the values of this work.

2. Related Work

SSL has been well studied for the classification task. We discuss some works which use deep networks since our target is to address the collapsing problem confronted by deep learning methods. Please refer to other surveys such as [36] for a more comprehensive review. Pseudo labeling [20, 26, 42, 43] is commonly used in SSL. The basic idea is to first learn an initial model on labeled images and then apply it to unlabeled images to estimate pseudo labels. The images with confident pseudo labels are added to the labeled dataset. Finally, it trains a stronger classifier on the extended dataset in a supervised way. However, the performance is limited by that of the initial classifier which is learned on only few labels. Iterative training alleviates the problem but the classifier is updated only once after it processes the whole dataset which is inefficient for large datasets. Besides, the selection criterion for data to be added to the labeled set is ad hoc for different tasks.

Some SSL methods [27, 19, 31, 2, 28] are consistency-based. For example, the Π model [19] keeps history predictions on the dataset and requires current predictions to be consistent with them. The approach is shown to be more tolerant to incorrect labels but is inefficient when learning large datasets since history predictions change only once per epoch. Tarvainen et al. [31] present the mean-teacher model in which the teacher is the moving average of the student which can be timely updated in every iteration. But their performance is limited because the two models tend to converge to the same point and stop further exploration. Some methods [25, 17] learn two different models by minimizing their prediction discrepancy. To avoid the case where the two models converge to the same point, they either learn from different initializations [17] or add view difference constraints [25]. Besides, there are some works that avoid collapsing without negative sample in self-supervised learning [10, 7, 44], but their objective functions and optimized variables are different from ours. The BYOL uses the Exponential Moving Average (EMA) strategy [10], which does not prevent collapsing in our experiments. The SimSiam shows that the stop-gradient plays an essential role [7] but using it alone without our easy-hard augmentation strategy also cannot avoid collapsing.

The above works were not been evaluated for pose estimation and we find they all encounter the collapsing problem when applied to the task. The contribution of this work lies in identifying and studying the collapsing problem and presenting a simple solution to avoid it such that the existing SSL methods can be used for pose estimation. In addition, we will extend some representative works to the human pose estimation task and provide a rigorous evaluation of their performance. This has empirical values to the community. We will release our code and models hoping it can facilitate research along this direction.
3. The Method

The task of 2D pose estimation aims to detect locations of $K$ body joints in an image $I$. Since [33], nearly all methods transform the problem to estimating $K$ Gaussian heatmaps $H$ where each heatmap encodes the probability of a joint at a location in $I$. Inference, each joint can be estimated to be at the location with the largest value in the corresponding heatmap. Denote the labeled and unlabeled training sets as $\mathcal{L} = \{(I^i, H^i)\}_{i=1}^N$ and $\mathcal{U} = \{I^u\}_{u=1}^M$, respectively. For supervised training of the pose estimation network $f$, we minimize the MSE loss between the estimated and ground-truth heatmaps:

$$L_s = \mathbb{E}_{I \in \mathcal{L}} \| f(I_{\eta}, \theta) - H_{\eta}\|^2,$$  \hspace{1cm} (1)

where $I_{\eta} = T(I, \eta)$ represents an augmentation of $I$ and $\eta$ represents augmentation parameter. $H_{\eta} = T(H, \eta)$ represents the corresponding heatmap and $\theta$ represents the network parameters.

3.1. Unsupervised Learning via Consistency

The network $f$ also learns about unlabeled images via consistency loss. For each unlabeled image $I$, it composes two augmentations $I_{\eta}$ and $I_{\eta'}$ and minimizes the MSE loss between the heatmap predictions:

$$L_u = \mathbb{E}_{I \in \mathcal{U}} \| f(I_{\eta}, \theta) - f(I_{\eta'}, \theta')\|^2.$$ \hspace{1cm} (2)

The network parameters $\theta$ and $\theta'$ can be either identical or different. For example, in [31], $\theta'$ is the exponential moving average (EMA) of $\theta$. We will evaluate both choices in our experiments. It is worth noting that both $\theta$ and $\theta'$ are changing during training. In contrast, the teacher network of the pseudo labeling methods is fixed so it does not suffer from collapsing. The parameters $\eta$ and $\eta'$ are usually randomly sampled at each training step. It is worth noting that $\eta$ and $\eta'$ are usually sampled from the same distribution without discrimination [2, 19, 31].

Figure 2. **Left:** the standard consistency-based method minimizes the distance between the predictions of the two augmentations (red and blue points). Since many pixels have low response (close to background), few high response pixels (e.g., the red point) tend to be gradually pushed to the background class. **Right:** In our method, more accurate predictions of easy augmentation pull those on hard augmentation, which avoids high response samples being pulled to the background class.

Figure 3. **Top:** results of the standard consistency-based method. Average heatmap response increases steadily for labeled images which is as expected. But for unlabeled images, it decreases to zero which suggests that collapsing occurs. The estimation accuracy on the validation dataset also decreases to 0.9%. **Below:** the results of our approach.

3.2. The Collapsing Problem

We try to train a model by adding the two loss functions: $L = L_s + \lambda L_u$, with $\lambda = 1$. Each batch of training data consists of equal number of images from $\mathcal{L}$ and $\mathcal{U}$. We use affine augmentation [30, 41] for $\eta$ and $\eta'$. We use identical weights for $\theta$ and $\theta'$ and use 1K labels. Within a few iterations of training, the network begins to predict all pixels of unlabeled images as background as shown in Figure 3 (top). The maximum value in a heatmap is used to represent its heatmap response and we find the average response on labeled images increases steadily which is as expected. However, the average response on unlabeled images decreases significantly and the accuracy on validation images is very low. Decreasing $\lambda$ does not solve the problem. It only slows down the collapsing process. So we set $\lambda = 1$ for the rest of our experiments. Some one may think it is over-fitting to the small labeled dataset. However, increasing labels to $118K$ does not fully solve the problem. The response on unlabeled images still gradually decreases. The accuracy is higher than the case with 1K labels but it is still worse than the initial supervised model. We also tried to use strong augmentation methods such as Rand Augmentation [8] to labeled images or unlabeled images but none of them can fully address the collapsing problem.

Collapsing occurs because the consistency regularization requires the model to satisfy the smoothness assumption [6, 36] where an image and its augmentation should have similar predictions. Thereby, the decision boundary would be pushed to low-density region. In fact, due to the imbalance in data, decision boundary often skews into the areas of minor class which is sparse globally as shown in Figure 4. This is also observed in [14]. As a result, a growing number of pixels are mis-classified as background.
3.3. Avoid Collapsing

The naïve implementation of the consistency regularization draws two samples to their middle point so more data are becoming closer to the decision boundary (see Figure 4.B). As a result, the decision boundary is pushed away from the high density areas of the dominant class and may skew into the areas of minor class. In contrast, our approach drives the less accurate predictions which are close to the decision boundary to the direction of more accurate predictions. In this case, the decision boundary is less likely to be incorrectly formed.

To achieve the goal, we present a paired easy-hard image augmentation strategy. For an unlabeled image $I$, it obtains two augmented images $I_e$ and $I_h$ by applying an easy and hard augmentation $T_e$ and $T_h$, respectively:

$$I_e = T_e(I) = T(I, \eta_e) \quad \text{and} \quad I_h = T_h(I) = T(I, \eta_h).$$ (3)

Where $T_e$ is regarded as an easier augmentation method than $T_h$ only when the network obtains better average accuracy on a dataset under perturbation $T_e$ than under $T_h$. We feed the two augmented images to the network and let the predictions of $I_e$ to teach the predictions of $I_h$:

$$L_{e,h} = \mathbb{E}_{I \in U} \left| |f(I_e, \theta) - f(I_h, \theta)| \right|^2.$$ (4)

For the sake of simplicity, we call $f(I_e, \theta')$ and $f(I_h, \theta)$ as teacher and student signals, respectively. Note that the gradients are propagated through only the student path. This is the key to avoid collapsing. This can be done by calling the detach operator on the teacher signals before computing the loss. Removing the detach operator leads to collapsing regardless of augmentations.

4. Implementation Details

4.1. Easy-Hard Augmentation

Affine Transformation is commonly used in 2D pose estimation which randomly scales and rotates an image. Affine transformation changes keypoint locations for pose estimation which are equivariant to the transformation [47, 32]. Let $T(\cdot)$ be an affine transformation and $f(\cdot)$ be the network to estimate heatmaps from images. Then the loss function can be computed as:

$$L = \mathbb{E}_{I \in U} \left| |f(T(I)) - f(T(I))| \right|^2.$$ (5)

It can be extended to map the heatmaps of the same image under different affine augmentations which allows us to compute the consistency loss.

We find that a pose estimator achieves very different performances on the same dataset if we apply affine transformation of different strengths to perturb the testing images. Figure 5 shows some typical results. For example, when we randomly sample rotation angles from $[-30^\circ, 30^\circ]$ and scale factors from $[0.75, 1.25]$ (denoted as “Affine 30”) for affine transformation to perturb testing images, the Average Precision (AP) on the dataset is about 63.7%. But when we sample from a larger range of $[-60^\circ, 60^\circ]$ and $[0.5, 1.5]$, respectively, AP decreases notably to 46.8%.

The above finding motivates us that we can compose easy-hard augmentation pairs by adapting the ranges of rotation and scaling. Figure 5 shows some easy-hard augmentation choices that are able to prevent the model from collapsing. It is worth noting that “Affine 60” can be regarded as a hard augmentation compared to “Affine 30”, but it can also be regarded as an easy augmentation when compared to a stronger method “JC 5” which we will introduce in the next section. It suggests that it is the gap between the two methods that matters.

Note that the augmentation strategies generalize well across datasets, which means that we need not to repeat the experimentation. In our experiments, the augmentations are determined based on 1K images sampled from COCO, and are applied to the rest datasets.
Joint Cutout Although affine-based augmentation already avoids collapsing, we find using harder augmentation for $T_h$ improves the accuracy ($T_e$ still uses easy affine augmentation). Inspired by cutout [9] and keypoint masking [16], we introduce a new method Joint Cutout to simulate occlusion. For each image (with easy augmentation applied), we first estimate coarse locations of keypoints using the model we are trying to train. Then we randomly sample a number of detected keypoints and mask their surrounding regions as illustrated in Figure 6. To avoid over-fitting to the masks, the center locations and sizes of masking regions are randomly perturbed. The method improves accuracy by a notable margin especially when the number of labeled images is small.

4.2. Learning Dual Networks

The previous SSL methods [31, 19, 28, 2] often learn a single network where the teacher’s parameters are either the same as the student’s or its exponential moving average. So the teacher and student networks are coupled which limits their performance [17]. The recent method [17] learns two independent networks to solve the problem. In this section, we briefly introduce how to apply easy-hard augmentation to it. For a training image $I$, we generate an easy and a hard augmentation denoted as $I_e$ and $I_h$, respectively. Then we feed them to both networks $f_h$ and $f_e$ and obtain four stream heatmap predictions:

$$
H_{θ,e} = f(I_e, θ) \quad \text{and} \quad H_{θ,h} = f(I_h, θ),
$$

$$
H_{ξ,e} = f(I_e, ξ) \quad \text{and} \quad H_{ξ,h} = f(I_h, ξ).
$$

We know that $H_{θ,e}$ and $H_{ξ,h}$ are similar up to a known transformation $T_e \rightarrow h$. Similarly, $H_{θ,h}$ is also similar to $H_{ξ,e}$ up to the same transformation. We train the networks by minimizing two consistency loss items:

$$
θ^* = \arg \min_θ \|H_{θ,h} - T_{e \rightarrow h}(H_{ξ,e})\|^2,
$$

$$
ξ^* = \arg \min_ξ \|H_{ξ,h} - T_{e \rightarrow h}(H_{θ,e})\|^2. \quad (7)
$$

We only pass the gradient back through the hard example to avoid collapsing. It means that one consistency loss item is used to optimize a single network at each time. Take the first formula in Eq. (7) as an example, $H_{ξ,e}$ estimated by $f_e$ is treated as a teacher to update $f_h$. In this case, we do not update $f_h$ because $H_{θ,h}$ is usually too noisy to be used as supervision. Subsequently, we update $f_e$ according to the second formula in Eq. (7). The two symmetrical loss items are combined so that the two networks can guide each other and be optimized together. The performances of the two networks are very close in the end and we report their average accuracy. Note that in inference, the model has the same number of parameters and running speed as the supervised model.

5. Baselines and Our Methods

We first introduce several baselines by modifying some representative SSL classifiers for pose estimation including both Pseudo labeling methods and consistency-based ones, and numerically compare them to our approach.

PseudoPose It is modified from pseudo labeling methods [20, 26, 42]. We first train a teacher model $f_t$ with labeled images. Then $f_t$ is fixed and we apply it to unlabeled images to obtain pseudo heatmaps. We train an ultimate model $f$ by minimizing the Mean squared error (MSE) loss on the combined set:

$$
L = \sum_{I \in L} \|H_e \leftarrow (f_t(I_e)) - f(I_h)\|^2 + \sum_{I \in U} \|H_e - f(I_e)\|^2,
$$

where $H_e$ is the ground-truth heatmap. Note that we use the same augmentation methods as ours for fair comparison.

DataDistill [26] It is also a pseudo labeling method. It differs from PseudoPose in that it sums the heatmaps estimated for multiple different augmentations of an image, obtains the keypoint locations, and re-generates a pseudo heatmap with Gaussian shape for supervision.

Ours (Single) It is a consistency-based method in which $θ$ and $θ'$ are identical. On labeled images, it performs supervised learning with the ground-truth heatmaps. For each unlabeled image, it minimizes the discrepancy between the two estimated heatmaps of the easy and hard augmented images. It differs from PseudoPose in that $f_t$ is not fixed. In fact, it is $f$ which is learned in semi-supervised learning.

Ours (Dual) The method is similar to “Ours (Single)” except that it learns dual networks as discussed in section 4.2. We also apply our proposed “easy-hard” augmentation method to this approach to avoid collapsing.
6. Experiment

6.1. Datasets, Metrics and Details

COCO Keypoint [22] It has four subsets of TRAIN, VAL, TEST-DEV and TEST-CHALLENGE. There are 123K WILD unlabeled images. To evaluate our method when different numbers of labels are used, we construct four training sets by randomly selecting 1K, 5K, 10K and 20K person instances from TRAIN, respectively. The unlabeled set consists of the rest of images from TRAIN unless specified. In some experiments, we use the whole TRAIN as the labeled set and WILD as the unlabeled one. We report the mean AP over 10 OKS thresholds as the main metric following [22]. The input image size is 256 × 192.

| Methods                  | Aug. | 1K  | 5K  | 10K | All   |
|--------------------------|------|-----|-----|-----|-------|
| Supervised [41]          | A    | 31.5| 46.4| 51.1| 67.1  |
| PseudoPose               | A    | 37.2| 50.9| 56.0| —     |
| DataDistill [26]         | A    | 37.6| 51.6| 56.6| —     |
| Ours (Single)            | A    | 38.5| 50.5| 55.4| —     |
| Ours (Dual)              | A    | 41.5| 54.8| 58.7| —     |
| Ours (Single) +JC        | A    | 42.1| 52.3| 57.3| —     |
| Ours (Dual) +RA          | A    | 43.7| 55.4| 59.3| —     |
| Ours (Dual) +JC          | A    | 44.6| 55.6| 59.6| —     |

MPII Dataset [1] It has about 25K images with 40K annotated person instances. Since labels are not provided for the test set, we conduct ablation study on the validation set which consists of 3K instances. We use the training set as the labeled set and the AI Challenger dataset [40] as the unlabeled set, which has 210K images with 370K person instances. The metric of PCKh@0.5 [1] is reported. The size of the input image is 256 × 256 following [1].

H36M Dataset [15] We use subjects S1, S5, S6, S7 and S8 for training, and S9, S11 for testing. The 2D pose accuracy is measured by Joint Detection Rate (JDR). And the Mean Per Joint Position Error (MPJPE) is used as the main metric in 3D pose estimation.

Implementation Details We use SimpleBaseline [41] to estimate heatmaps and ResNet18 [12] as its backbone. But our approach is general and can be applied to other pose estimators as shown in Table 4. On the validation set, we use the ground truth boxes and do not flip images for all methods. We train the models for 100 epochs. We use Adam [18] optimizer with the initial learning rate of 1e−3. It drops to 1e−4 and 1e−5 at 70 and 90 epochs, respectively.

Supervised [41] HRNet w48 39.2 57.7 63.7

Ours (Dual) HRNet w48 50.9 64.3 67.9

Ours (Dual) Ours (Dual) 48.7 59.4 62.5

6.2. Ablative Study

Easy-Hard Augmentation We first study the relationship between augmentation methods and collapsing. As shown in Figure 7 (a), when we use easy augmentations for both $T_e$ and $T_h$, the average response gradually decreases to zero for unlabeled images meaning collapsing occurs. This is because there is no accuracy gap between teacher and student signals. We also get degenerated results when we use hard augmentation for $T_e$ (see sub-figures c and d) for the same reason. In contrast, the training becomes normal when we use easy-hard augmentation strategy. In this case, the teacher and student models have sufficient gap. We have similar observation when we either learn a single model or dual models.

Baseline SSL Methods Table 1 shows the results of different SSL methods. Supervised training with a small number of labels gets worst results which validates the values of unlabeled images. The gap is larger when there are fewer labels. DataDistill [26] achieves slightly better accuracy than PseudoPose since it ensembles multiple network output to obtain more reliable pseudo labels. The proposed consistency-based method “Ours (Dual)” get better results than the pseudo labeling methods.

We also study the impact of augmentation methods for $T_h$. We can see from Table 1 (bottom) that applying harder augmentation methods such as RandAug (RA) and Joint Cutout (JC) notably improves the results especially when the number of labeled images is small. In particular, Joint Cutout achieves consistently better mean AP scores than RandAug. It is known that the most common mistake in pose estimation is the confusion between left and right joints. As shown in Figure 6, Joint Cutout is effective in increasing the level of confusion and drives the models to learn more discriminative features. We use Joint Cutout as the default augmentation for the rest of the paper. It is worth noting that applying hard augmentation to [41] in supervised training actually decreases AP when there are 1K labels and slightly increases AP from 51.1% to 52.1% when there are 10K labels.

Table 1. AP of different methods on COCO when different numbers of labels are used.

Table 2. The effects of using different network structures for the two models $f_o$ and $f_c$ on COCO. We report AP when different numbers of labels are used.
### Network Structures

We evaluate the effect of using different networks in Table 2. We can see that when we use ResNet18 and HRNet, the performance of ResNet18 improves by a large margin (41.5% vs. 48.7%) compared to the case of using ResNet18 for both networks. This is mainly because HRNet can provide more accurate supervision for ResNet18 which notably boosts its performance. The results suggest that, even when our target is to learn a lightweight model for fast inference, we can still learn it together with a large model which will notably improve the accuracy of the lightweight model.

### 6.3. Failed Attempts

We present some failed attempts to avoid collapsing. The first is to balance foreground/background pixels by class re-weighting. Since we do not have labels, we assign larger weights for pixels with larger heatmap predictions since they are more likely from foreground. We tried several weight functions but collapsing still occurs because we do not have ground-truth labels (see Figure 8 (C)). The second approach uses confident predictions to teach the network. If the maximum response of a heatmap (of teacher) is larger than a threshold, we use it as supervision. Otherwise, we do not use it in loss computation. When the threshold is small, the performance is much worse than the initial supervised model (see (A1)). When the threshold is large, very few pixels are involved in training and the performance is barely improved (see (A4)). The third approach stabilizes training using Mean Teacher [31] in which the teacher is the exponential moving average of the student $\theta' \leftarrow \alpha \theta' + (1 - \alpha) \theta$. We set $\alpha$ to be 0.99 and 0.999, respectively. The performance is worse than the initial supervised model which does not use unlabeled images (see (B1-B2)).

### 6.4. Performance with Many Labels

We use COCO TRAIN and WILD as labeled and unlabeled datasets, respectively. The results on the VAL set are in Table 3. Our approach consistently outperforms the initial supervised model. It suggests that even when we have access to many labels, it still gets decent improvement with unlabeled images. We also test our approach in a more realistic setting where labeled and unlabeled images are from different datasets of MPII and AIC, respectively. Table 5 shows the results on the test set of MPII. Our approach outperforms all other methods. The experiment validates the values of using unlabeled images. The last two methods use extra labels and larger image sizes.

Table 4 shows the results of the state-of-the-art methods on the COCO test-dev dataset. We supplement them with our approach to learn about unlabeled images from the COCO WILD dataset. We can see that our approach consistently improves the performance although the performance of the original methods are already very high.
SSL can also be used for unsupervised domain adaptation to learn about unlabeled images from a new domain. To that end, we evaluate different methods by trying to adapt the model learned on the MPII dataset to the H36M dataset. We first estimate the model learned on the MPII dataset to the H36M dataset. The MPII is used as labeled set and H36M is used as unlabeled set. No labels from H36M are used in training. Table 6 shows the results. Directly using the model trained on H36M achieves notably higher 2D pose estimation accuracy and lower 3D pose error than the model pre-trained only on the labeled dataset MPII and finetuned on H36M.

In this work, we present the first systematic study of semi-supervised 2D pose estimation. In particular, we first identify and discuss the collapsing problem in consistency based methods. Then we present a simple yet effective approach to solve the problem. We conduct extensive experiments to validate the effectiveness of our approach and show that it can benefit many different application scenarios. We released our code and models hoping to inspire more research in this direction.

**Acknowledgement**

This work was supported in part by MOST-2018AAA0102004, NSFC-61625201, 61527804 and DFG TRR169 / NSFC Major International Collaboration Project "Crossmodal Learning".
References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3686–3693, 2014. 2, 6

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In Advances in Neural Information Processing Systems, pages 5049–5059, 2019. 1, 2, 3, 5

[3] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. In European Conference on Computer Vision, pages 606–622, 2020. 8

[4] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. In IEEE International Conference on Automatic Face and Gesture Recognition, pages 8–15. IEEE, 2020. 8

[5] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Itendra Malik. Human pose estimation with iterative error feedback. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4733–4742, 2016. 1

[6] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. Semi-Supervised Learning. The MIT Press, 1st edition, 2010. 3

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15750–15758, 2021. 2

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 702–703, 2020. 3, 4

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552, 2017. 5

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Ghsashlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, Le. Randaugment: Practical automated data augmenta-

[11] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003. 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016. 6

[13] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5700–5709, 2020. 8

[14] Minsung Hyun, Jisoo Jeong, and Nojun Kwak. Class-imbalance semi-supervised learning. arXiv preprint arXiv:2002.06815, 2020. 2, 3

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1325–1339, 2014. 2, 6, 8

[16] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In European Conference on Computer Vision, pages 713–728, 2018. 5, 8

[17] 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 Proceedings of the IEEE International Conference on Computer Vision, pages 6728–6736, 2019. 2, 5

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015. 6

[19] Samuli Laine and Timo Aila. A self-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, volume 3, 2013. 2, 5

[20] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, volume 3, 2013. 2, 5

[21] Wenbo Li, Zhicheng Wang, Binya Yin, Qixiang Peng, Yuming Du, Tianzai Xiao, Gang Yu, Hongtiao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148, 2019. 8

[22] 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 European Conference on Computer Vision, pages 740–755. Springer, 2014. 2, 6

[23] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Hai Ci, and Yizhou Wang. Context modeling in 3d human pose estimation: A unified perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6238–6247, 2021. 1

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In European Conference on Computer Vision, pages 483–499. 2016. 1, 8

[25] Sijuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In European Conference on Computer Vision, pages 135–152. 2018. 2

[26] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4119–4128, 2018. 1, 2, 3, 5, 6
[27] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, page 1171–1179, Red Hook, NY, USA, 2016. Curran Associates Inc. 1, 2

[28] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In Neural Information Processing Systems, 2020. 1, 2, 5

[29] Zhihui Su, Ming Ye, Guohui Zhang, Lei Dai, and Jianda Sheng. Cascade feature aggregation for human pose estimation. arXiv preprint arXiv:1902.07837, 2019. 8

[30] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5693–5703, 2019. 1, 3, 7, 8

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in neural information processing systems, pages 1195–1204, 2017. 1, 2, 3, 5, 7

[32] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In Proceedings of the IEEE international conference on computer vision, pages 5916–5925, 2017. 4

[33] 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 Advances in neural information processing systems, pages 1799–1807, 2014. 1, 3

[34] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1653–1660, 2014. 1

[35] Han Yue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In ECCV, pages 197–212. Springer, 2020. 1

[36] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. Machine Learning, 109(2):373–440, 2020. 2, 3

[37] Chunyu Wang, Yizhou Wang, and Alan L Yuille. An approach to pose-based action recognition. In CVPR, pages 915–922, 2013. 1

[38] Chunyu Wang, Yizhou Wang, and Alan L Yuille. Mining 3d key-pose-motifs for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2639–2647, 2016. 1

[39] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4724–4732, 2016. 1

[40] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, and Y. Wang. Large-scale datasets for going deeper in image understanding. In IEEE International Conference on Multimedia and Expo (ICME), pages 1480–1485, 2019. 6

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In European Conference on Computer Vision, pages 466–481, 2018. 1, 3, 6, 7, 8

[42] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10687–10698, 2020. 1, 2, 5

[43] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In 33rd annual meeting of the association for computational linguistics, pages 189–196, 1995. 2

[44] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, pages 12310–12320, 2021. 2

[45] Feng Zhang, Xiutian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7093–7102, 2020. 8

[46] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia. Human pose estimation with spatial contextual information. arXiv preprint arXiv:1901.01760, 2019. 8

[47] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2694–2703, 2018. 4

[48] Zhe Zhang, Chunyu Wang, Wenhui Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In CVPR, pages 2200–2209, 2020. 1