Inter-image Contrastive Consistency for Multi-Person Pose Estimation

Xixia Xu[1†], Yingguo Gao[2], Xingjia Pan[2], Ke Yan[2*], Xiaoyu Chen[1], Qi Zou[1*]

[1] Beijing Key Laboratory of Traffic Data Analysis and Mining,
Beijing Jiaotong University, Beijing, China
[2] Tencent Youtu Lab, Shanghai, China
{19112036, 22110103, qzou}@bjtu.edu.cn, {yingguogao, kerwinyan}@tencent.com, xjia.pan@gmail.com

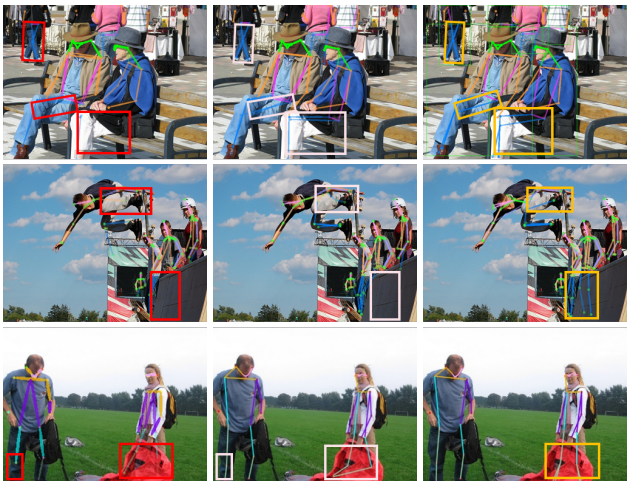
Abstract

Multi-person pose estimation (MPPE) has achieved impressive progress in recent years. However, due to the large variance of appearances among images or occlusions, the model can hardly learn consistent patterns enough, which leads to severe location jitter and missing issues. In this study, we propose a novel framework, termed **I**nter-image **Con**trastive consistency (ICON), to strengthen the keypoint consistency among images for MPPE. Concretely, we consider two-fold consistency constraints, which include single keypoint contrastive consistency (SKCC) and pair relation contrastive consistency (PRCC). The SKCC learns to strengthen the consistency of individual keypoints across images in the same category to improve the category-specific robustness. Only with SKCC, the model can effectively reduce location errors caused by large appearance variations, but remains challenging with extreme postures (*e.g.*, occlusions) due to lack of relational guidance. Therefore, PRCC is proposed to strengthen the consistency of pair-wise joint relation between images to preserve the instructive relation. Cooperating with SKCC, PRCC further improves structure aware robustness in handling extreme postures. Extensive experiments on kinds of architectures across three datasets (*i.e.*, MS-COCO, MPII, CrowdPose) show the proposed ICON achieves substantial improvements over baselines. Furthermore, ICON under the semi-supervised setup can obtain comparable results with the fully-supervised methods using only 30% labeled data.

Introduction

Multi-person pose estimation (MPPE) aims to locate human keypoints (*e.g.*, neck, wrist) for each person in an image, which is a basic yet challenging task in computer vision. Benefiting from the development of deep learning(He et al. 2016), CNN-based methods made significant progress. However, the ability in handling keypoints with large appearance variations or extreme postures is still limited.

Model training is a procedure from special to general and aims to learn consistent representations. Large variation of appearances or extreme postures make it hard to learn consistent patterns across images for accurate estimation. To address these issues, some existing methods(Chen et al. 2018; Su et al. 2019b; Sun et al. 2019) design elaborate architectures to learn more precise features or adopt the attention mechanism(Su et al. 2019a; Chu et al. 2017) to capture discriminative features for accurate result. Although these methods improve feature representations, they still suffer from the inconsistent joint representations between images due to lack of explicit consistency constraints.



(a) Baseline (b) +SKCC (c) +SKCC+PRCC

Figure 1: Challenges existed in MPPE: (a) the results with the current intra-image baseline; (b) the SKCC results based on baseline; (c) the final results are refined by PRCC.

To strengthen the representation consistency, one line of researches focus on the graph-based methods (Wang et al. 2020; Qiu et al. 2020a,b) to model the structural relations (*e.g.,* wrist is adjacent to elbow). GPCNN(Wang et al. 2020) proposes a refinement module to capture the joint context, but the relations with less semantic correlations may mislead the model. To dynamically model rich relations among keypoints, DGCN(Qiu et al. 2020b) propose a dynamic graph to tolerate the large variations of human pose. Either empirically picking out edges or dynamically learning relations from the graph, these methods consider most of relations, which inevitably involve redundant relations with dedicate module design. Besides, only considering intra-image relations also limits further improvement for MPPE.

[†]Works done while interning at Tencent Youtu Lab.
[*]Corresponding author.

Recently, another line of methods to retain the explicit semantic consistency, the contrastive learning (CL) (Chen et al. 2020; He et al. 2020) emerges. It has shown great potential in improving the representations via discriminating similar image pairs (constructed from different augmentations of the same images) from dissimilar image pairs. Some researchers also applied the idea of CL in the dense pixel-level prediction tasks like segmentation (Jiang et al. 2021; Liu et al. 2021), but it remains unexploited for MPPE. Human poses have inherent structure while these methods only consider the pixel-level or sample-wise similarities, ignoring semantic relation between keypoints across images, while such relation cues can greatly help model correct the structural errors for extreme postures (Fig. 1).

In this study, we propose an **I**nter-image **CON**trastive consistency **(ICON)** strategy, making the first attempt to strengthen the inter-image consistency for MPPE. Specifically, ICON includes the *Single Keypoint Contrastive Consistency* **(SKCC)** and *Pair Relation Contrastive Consistency* **(PRCC)**, as shown in Fig. 2. The SKCC considers the consistency of individual keypoint within the same category between images. Only with SKCC, the model can effectively alleviate the location issues caused by large variations of appearance, but it still remains challenging when encounters extreme postures (Fig. 1 (b)). In contrast to existing graph-based methods that only considering intra-image joint relations, we instead construct the pair relations of keypoints and propose the PRCC to strengthen the consistency between keypoint pairs across images for preserving inter-joint semantic relations. Cooperating with SKCC, the PRCC further empowers the model in handling extreme postures (Fig. 1 (c)) with the learned relation knowledge. We conduct extensive experiments on kinds of architectures (*i.e.*, SimpleBaseline, HRNet) and backbones (*i.e.*, ResNet, HRNet), and obtain significant improvements across three datasets (*i.e.*, MS-COCO, MPII and CrowdPose). Moreover, we apply the proposed approach under the semi-supervised setup and our method can obtain comparable results with the fully-supervised method using only 30% labeled data.

In a nutshell, the contributions of this work are as follows:

- We are creatively to explore the inter-image contrastive consistency in 2D MPPE to learn robust keypoint representations.
- We propose the single keypoint and pair relation contrastive consistencies to effectively strengthen the inter-image consistent keypoint features and relation features, significantly alleviating the location issues.
- The ICON achieves substantial and consistent improvements over previous methods on three challenging MPPE datasets under both fully and semi-supervised setup.

Related Work

Multi-Person Pose Estimation

Recently, researchers have made great efforts(Chou, Chien, and Chen 2018; He et al. 2017; Ke et al. 2018; Bin et al. 2020; Tang, Yu, and Wu 2018; Ning, Zhang, and He 2018) to make progress as the development of the deep learning. Two mainstream methods are prevalent in MPPE including bottom-up(Cheng et al. 2020; Wei et al. 2016) and top-down(Bulat et al. 2020; Newell, Yang, and Deng 2016; Yang et al. 2017; Tang et al. 2018). The former directly estimates all keypoints and assigns them into persons. The latter detects the human bounding boxes and then locates the keypoints within each box. Besides, graph-based methods(Jin et al. 2020; Qiu et al. 2020b; Wang et al. 2020) model the joint relation context and also achieve great performance. The related semi-supervised work(Moskvyak et al. 2021) enforces similar features for the same keypoint class across dataset but it needs an additional keypoint classifier and involves the attention-based features.

Our method follows the top-down pipeline and differs from previous works that improve estimation quality via learning better features or keypoint relations inside the image. We propose a novel perspective to alleviate the localization errors via mining the inter-image contrastive consistency of keypoint representations in a simple manner.

Contrastive Learning

Recently, contrastive learning contributes to the success in self-supervised methods(Chen et al. 2020; He et al. 2020; Khosla et al. 2020), by learning a similarity function to bring augmented views of the same data closer in feature space, and pushing augmented views of different data apart. Most works learn similarity based on global representations of the different views to assist the classification but show limited supervision for dense prediction tasks. Dense representations rely on pixel-level features, naturally providing extra supervision capturing intrinsic spatial structure and fine-grained pixel correspondence. Though some concurrent works also address contrastive learning in dense image prediction(Chaitanya et al. 2020; Xie et al. 2021b), the ideas are quite different. (1) They typically regard contrastive learning as a pre-training step for embedding, while we propose two complementary consistency constraints jointly trained with the main task. (2) They compute the pixel-level contrast from augmented versions of the same image, ignoring the contrastive learning across diverse images. (3) Most importantly, we design consistency constraint of category-wise keypoint relations, which is never considered before.

Methodology

Approach Overview

Let (X, Y) be a training dataset with training images $x \in X$ and their corresponding K types of keypoint location labels $y \in Y$, where y can be acquired from the original dataset (supervised setting) or generated as pseudo-labels (semi-supervised setting). This pose estimator can be divided into two parts: an encoder network $\phi : X \rightarrow \varphi$, and a decoder regression head $\psi_r : \varphi \rightarrow \hat{Y}$. As shown in Fig. 2, we constrain pixel-level contrastive consistency on keypoint representations. This is done by attaching a representation projection head f_{proj} on top of the encoder ϕ, parallel to the regression head. Specifically, the encoder ϕ (*e.g.*, ResNet) outputs a dense feature map $F \in \mathcal{R}^{H \times W \times D}$, and then are mapped into a higher m-dimensional feature

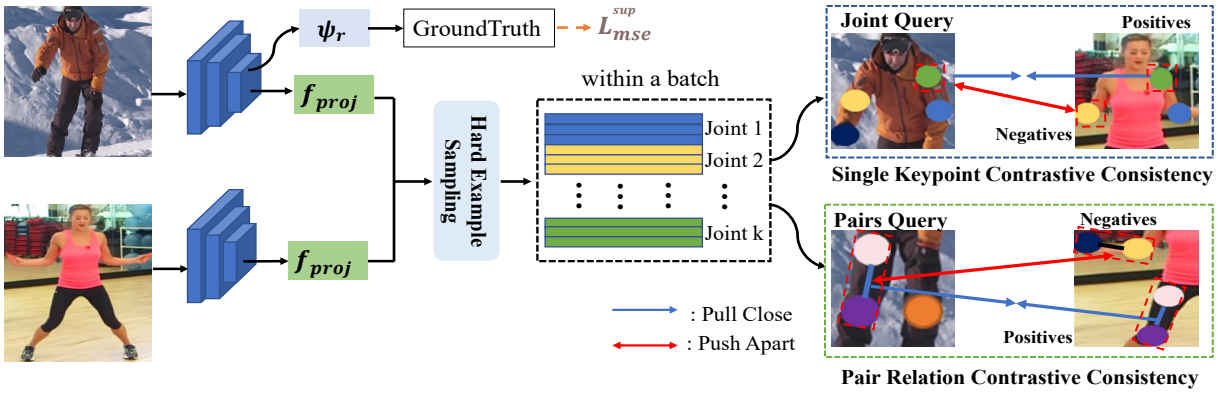


Figure 2: Overview of the ICON based MPPE architecture. Input images are firstly fed into a human detector to obtain the detected instances. Then, each instance is handled by the pose estimator trained with the proposed ICON constraint.

$F_k \in \mathcal{R}^m$, from which each keypoint embedding f_k (the colored feature strip "within a batch" box in Fig. 2) can be derived according to the coordinate groundtruth The representation head is only applied during training to guide the regressor considering the ICON constraint, and is removed during inference.

Single Keypoint Contrastive Consistency

The distances of keypoint features with the same category tend to be close in feature space. Although a regression loss can help model to learn the discriminative features for prediction, the learned features of keypoints with the same category across images still show significant intra-class variations caused by large variance of appearances as in Fig. 5 (a). Thus, we argue that by keeping the consistency between features of keypoints in the same category, *i.e.*, strengthening in a contrastive manner, more precise locations can be located from the augmented keypoint features.

We propose SKCC to drive the consistency of features from different keypoints of the same category in a mini-batch N. N denotes the instance numbers within an batch. In specific, for a category-specific keypoint query f_k belonging to a certain instance with its category k, the positive samples are other pixels also belonging to the same category, while the negatives are the pixels belonging to other categories across N instances. The SKCC loss is depicted as:

$$\mathcal{L}_{sin} = \frac{1}{\mathcal{P}_i} \sum_{i^+ \in \mathcal{P}_i} -log \frac{exp(i \cdot i^+ / \tau)}{exp(i \cdot i^+ / \tau) + \sum_{i^- \in \mathcal{N}_i} exp(i \cdot i^- / \tau)}, \quad (1)$$

where i denotes f_k, $\mathcal{P}_i$ and $\mathcal{N}_i$ denote pixel embedding sets of the positive and negative samples, respectively. Note that all embeddings are L_2-normalized. Thus, the SKCC enables the model learn more precise and robust keypoint features by making the intra-class features consistent and pushing inter-class features separated.

Pair Relation Contrastive Consistency

We observe that only relying on the single keypoint features cannot well deal with the extreme postures such as occlusions. To this end, the higher-order relations among joints, which is instructive for inferring the missing locations and correcting structural errors (*i.e.*, occlusion, inversion) should be explored. Previous methods model the relation representations via graph, which only considers the joint correlations within an image as in Fig. 3 (a), and involves sophisticated modules and computation costs. In ours, we finally realize that it's efficient and effective to preserve inter-category joint relations via exploring the pair-wise relation consistency across instances. More analyses about 1) the relation modeling between graph-based methods and ours and 2) the higher-order keypoint (triplet or quaternion) relations explorations are discussed in Suppl. Material.

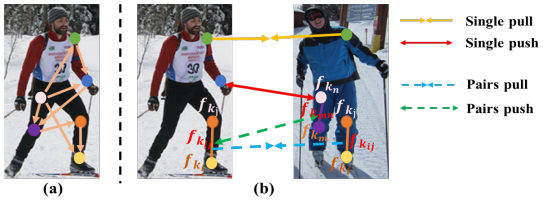


Figure 3: The difference between (a) intra-image feature learning and (b) inter-image keypoint feature exploration. The solid and dotted line denote the SKCC and PRCC.

Specifically, for each joint pairs belonging to an instance, we adopt $f_{k_{ij}}$ to denote the pair-wise relation features via considering the representations of the joint k_i and k_j (f_{k_i},f_{k_j}) as in Fig. 3 (b). Furthermore, considering the spatial distribution of joints, we also take the spatial relationship of whether the two joints are related into account and embed with a relation indicator $\mathbb{M}$ as:

$$\mathbb{M}_{k_{ij}} = \begin{cases} 1, & < k_i, k_j > adjacent, \\ 0, & otherwise. \end{cases} \quad (2)$$

Combined with this structural relation prior, we design the pair-wise relational feature $f_{k_{ij}}$ as,

$$f_{k_{ij}} = \mathbb{M}_{k_{ij}} \cdot (f_{k_i} + f_{k_j})/2. \quad (3)$$

To ensure the pair-wise relation consistency, we pull the pair relations of the same joint categories across images close and push the different ones apart sequentially. Concretely, for a category-specific pair-wise sampled query v

from $f_{k_{ij}}$, the positive samples are the representations belonging to the same category of joint pairs in other instances while the negatives are the features belonging to the joint pairs in other categories. The pair-wise loss is defined as:

$$\mathcal{L}_{pair} = \frac{1}{\mathcal{P}_v} \sum_{v^+ \in \mathcal{P}_v} -log \frac{exp(v \cdot v^+ / \tau)}{exp(v \cdot v^+ / \tau) + \sum_{v^- \in \mathcal{N}_v} exp(v \cdot v^- / \tau)}, \tag{4}$$

where $\mathcal{P}_v$ and $\mathcal{N}_v$ denote the pair-wise relation embedding sets of the positive and negative samples.

Hard Sampling on Queries and Negatives

Prior works(Kalantidis et al. 2020) find that the discriminating power of the training samples is crucial for contrastive learning. Considering our case, not all categories of joints need to be considered when learning consistency. For example, the categories that can be easily located contribute little to the optimization while cost huge memory and slow down the training. Thus, we design a hard sampling strategy to sample a sparse set of joint samples. Noted that when a category-specific queried joint (joint pair) in current image is decided, the positive ones are in the same category from other instances (within a batch), the rest of the joint (joint pair) categories are viewed as the negatives.

For PRCC, we construct the joint pairs according to the human skeleton spatial distribution. Concretely, we divide all joints into 8 pairs (joint-to-joint) and each pair consists of the adjacent joints in spatial (*i.e., lelbow* $\rightarrow$ *lwrist*). The specific sampling strategies are explored as follows.

• *ALL.* Each class of joint (joint pair) as the query and the other categories in other instances as negative ones.

• *Random.* The keypoint categories of the queries and the negatives are randomly selected and the number of the sampled queried categories adopt $6 \sim 9$, $3 \sim 5$ and negative categories adopt $5 \sim 10$, $4 \sim 8$ for SKCC and PRCC.

• *Hard Example Sampling.* We choose the hard examples (joints or joint pairs) with larger uncertainty (entropy of joint prediction value) and prediction loss as our queries. And we choose top-Q ($Q = 5$) as queried joint for SKCC and all joint pairs ($Q = 8$) for PRCC. Once a queried joint is selected, we calculate the scores of feature similarity between the query and all negative samples, then select top-H ($H = 6$ for SKCC and $H = 4$ for PRCC) the most similar categories as the negative ones with the support of the experimental results. The feature similarity function adopts the cosine distance. The hard sampling strategy enables the model focus more on the harder keypoint categories and delivers more accurate prediction.

ICON under Fully/Semi-supervised Setting

The consistency constraints can be easily applied in both fully and semi-supervised MPPE methods without changing the training pipeline. To incorporate ICON, we attach the constraints to the projected representation head directly. In the supervised setting, we adopt the MSE loss to measure the groundtruth P^* and prediction P. Combined with the constraint in Eq. 1, 4, the total loss is depicted as below:

$$\begin{aligned} \mathcal{L}_{ctrcon} &= \lambda_1 \mathcal{L}_{sin} + \lambda_2 \mathcal{L}_{pair}, \\ \mathcal{L}^{sup}_{mse} &= \sum_{I \in \mathcal{L}} MSE(P^*, P), \\ \mathcal{L}_{total} &= \mathcal{L}^{sup}_{mse} + \lambda_p \mathcal{L}_{ctrcon}. \end{aligned} \tag{5}$$

In the semi-supervised setup, we construct two semi-supervised MPPE baselines (pseudo-based methods(Lee et al. 2013; Radosavovic et al. 2018) to modify for MPPE and a recent consistency-based SSPose(Xie et al. 2021a)) to improve their performance by providing the inter-image consistency supervision. In the former, we firstly train a teacher model f_t with labeled images, and we apply f_t to unlabeled images to obtain reliable pseudo heatmaps whose predicted confidence is above a threshold γ. We train the model f by minimizing the MSE loss on the combined set:

$$\mathcal{L}^{semi}_{mse} = \sum_{I \in \mathcal{L}} MSE(P^*, f(I)) + \lambda_u \sum_{I \in \mathcal{U}} MSE(f_t(I), f(I)), \tag{6}$$

where $\mathcal{L}$ and $\mathcal{U}$ denote the labeled and unlabeled set. The λ_u reweights the contribution of unsupervised loss making sure the model cannot be dominated by gradients of uncertain pseudo-labels. The weighted total loss is as below:

$$\mathcal{L}_{total} = \mathcal{L}^{semi}_{mse} + \lambda_p \mathcal{L}_{ctrcon}, \tag{7}$$

where λ_p is the weight factor.

Experiments demonstrate that the regression loss and consistency constraint complement each other. The former lets the model learn spatial precise features for accurate estimation, while the latter helps to regularize the features with improved category-specific features and inter-joint relations through explicitly strengthening the inter-instance contrastive consistency of keypoint representation.

Experiments

Datasets and Experimental Setup

MS-COCO(Lin et al. 2014) follows (Sun et al. 2019).
MPII(Andriluka et al. 2014) includes $25K$ images with $40K$ objects, there are $12K$ objects for testing and the remaining for training. We use *PCKh* (head-normalized probability of correct keypoint) score as evaluation.
CrowdPose(Li et al. 2019) contains $20K$ images and $80K$ human instances which aims to promote performance in crowded cases and uses the same evaluation with COCO. It divides into three crowding levels by Crowd Index: *Easy* ($0 \sim 0.1$), *Medium* ($0.1 \sim 0.8$) and *Hard* ($0.8 \sim 1$).
SSL training set partition. We evaluate our method when different ratios of labeled instances are used on COCO and MPII. We choose $5\%, 10\%, 20\%, 30\%$ of the training samples as the labeled data and the remaining are unlabeled.

Implementation Details

We implement all experiments in PyTorch with 4 Tesla V100 GPUs. We adopt adopt HR-w32 and ResNet-50 as the default backbones, which are initialized with the pretrained ImageNet weight. The representation head is taken from the

feature of stage-3 in HRNet and res-4 in ResNet and the batch size is 16. Noted that only the visible keypoint categories are considered according to the labels for an instance, the occluded or unlabeled joints are ignored.

For COCO, the detected human boxes are resized to 256×192 and we trained for 210 epochs. The learning rate follows (Sun et al. 2019). For MPII, input size is 384×384 and the trained for 180 epochs. For CrowdPose, the training is similar with COCO and trained for 220 epochs. For data augmentation, we apply random flip with probability of 0.5, random rotation in $[-45°,+45°]$, random resize with $[0.65, 1.35]$ and half-body augmentations. The weighted factors in Eq. 5, 7 are set as: $\lambda_1 = 0.3$, $\lambda_2 = 0.7$, $\lambda_p = 1$, $\lambda_u = 0.1$. The $\tau = 0.07$, $\gamma = 0.75$.

Comparisons with the State-of-the-arts

On MPII. We evaluate the PCKh@0.5 score in Tab. 1. We find that, by equipping with cross-image consistency supervision, the performance of baseline enjoys consistent improvements $(1.0\%, 0.9\%)$ over ResNet, HRNet, especially for the hard joints (*i.e., wrist, knee and ankle*). The improvement mainly benefits from more precise keypoint features and relation guidance learnt by the consistency constraint.

Method	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Total
Wei	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell	98.2	96.3	91.2	87.2	89.8	87.4	83.6	90.9
Sun	98.1	96.2	91.2	87.2	89.8	87.4	84.1	91.0
Tang	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Ning	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Yang	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Su*	98.7	97.5	94.3	90.7	93.4	92.2	88.4	93.9
Bin*	**98.9**	**97.6**	**94.6**	**91.2**	93.1	**92.7**	89.1	**94.1**
Bulat*	98.8	97.5	94.4	**91.2**	**93.2**	92.2	**89.3**	**94.1**
SBN	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
+ICON	98.6	97.0	92.9	90.0	91.6	90.6	86.3	92.5
HRNet	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
+ICON	98.7	97.2	93.5	90.4	92.4	91.5	88.5	93.2

Table 1: Results comparisons on MPII *test* set. * means extra data and larger input size are used. The **bold** means the best.

On CrowdPose. To validate the effectiveness of the ICON even in crowd scenarios, we conduct experiments on CrowdPose. ICON strategy improves the mAP of base networks by solid margis (i.e., $+1.3\%$ for ResNet, $+0.5\%$ for OPECNet) in Tab. 2. We also report AP at different crowding levels (*M-medium, H-hard*). The improvements remain stable even the crowd index is high, which is impressive because the improvement on this complex dataset is so hard.

On MS-COCO. We combine ICON with the baselines and compare with the recent methods on COCO *test-dev* set in Tab. 3. ICON improves the performance of base networks by solid margins (*i.e.*, 1.6% for HRNet, 1.7% for ResNet)

Method	AP	AP_{50}	AP_{75}	AP_M	AP_H
Bottom-up methods					
OpenPose(Cao et al. 2019)	-	-	-	48.7	32.3
HrHRNet(Cheng et al. 2020)	67.6	87.4	72.6	68.1	58.9
DEKR(Geng et al. 2021)	67.3	86.4	72.2	68.1	58.7
SWAHR(Luo et al. 2021)	71.6	88.5	77.6	72.4	63.0
PETR(Shi et al. 2022)	**72.0**	**90.9**	**78.8**	**72.5**	**65.4**
Top-down methods					
SBN(Sun et al. 2017)	60.8	81.4	65.7	61.2	51.2
RMPE(Fang et al. 2017)	61.0	81.3	66.0	61.4	51.1
SPPE(Li et al. 2019)	66.0	84.2	71.5	66.3	57.4
OPECNet(Qiu et al. 2020a)	70.6	86.8	75.6	-	-
ICON+SBN	62.1	82.9	65.8	62.4	51.5
ICON+OPECNet	71.1	87.2	75.9	-	-

Table 2: Performance comparisons on CrowdPose *test* set.

and surpasses most methods. We can achieve best result under the same condition. It shows the learnt pair-wise joint relations and robust keypoint features are indeed effective.

Method	AP	AP_{50}	AP_{75}	AP_M	AP_L
Bottom-up methods					
HrHRNet(Cheng et al. 2020)	66.4	87.5	72.8	61.2	74.2
SWAHR(Luo et al. 2021)	67.9	88.9	74.5	62.4	75.5
SIMPLE(Zhang et al. 2021)	69.6	89.3	77.9	68.1	77.8
DEKR(Geng et al. 2021)	69.8	89.0	76.6	65.2	76.5
Top-down methods					
CPN(Chen et al. 2018)	73.0	91.7	80.9	69.5	78.1
CSANet(Su et al. 2019a)	74.5	91.7	82.1	71.2	80.2
UDP(Huang et al. 2020)	74.7	91.8	82.1	71.5	80.8
DGCN(Qiu et al. 2020b)	67.4	88.0	74.4	63.6	73.0
GPCNN(Wang et al. 2020)	76.4	92.5	83.8	72.9	**82.4**
OPECNet(Qiu et al. 2020a)	73.9	91.9	82.2	-	-
SBN(Sun et al. 2017)	71.9	91.4	80.1	68.9	77.4
+ICON(256×192)	73.6	91.6	81.0	70.2	79.8
SBN(Sun et al. 2017)	73.7	91.9	81.1	70.3	80.0
+ICON(384×288)	75.2	92.3	83.0	71.6	80.5
HRNet-w32(Sun et al. 2019)	73.5	92.2	82.0	70.4	79.0
+ICON(256×192)	75.1	92.3	82.7	71.9	80.7
HRNet-w32(Sun et al. 2019)	74.9	92.5	82.8	71.3	80.9
+ICON(384×288)	76.0	92.4	83.5	72.6	82.0
HRNet-w48(Sun et al. 2019)	74.2	92.4	82.4	70.9	79.7
+ICON(256×192)	75.4	92.4	83.2	72.0	81.4
HRNet-w48(Sun et al. 2019)	75.5	92.5	83.3	71.9	81.5
+ICON(384×288)	**76.5**	**92.5**	**83.9**	**73.0**	82.3

Table 3: Performance comparisons on COCO *test-dev* set.

Qualitative Results. Fig. 4 depicts qualitative comparisons of HRNet+Ours against HRNet over representative examples. As base model couldn't locate keypoints accurately for extreme postures (*i.e.*, self-occlusions) or in crowded scenes. These issues are well eased by ours benefiting from the robust keypoint features captured by SKCC and the preserved relation learned by PRCC.

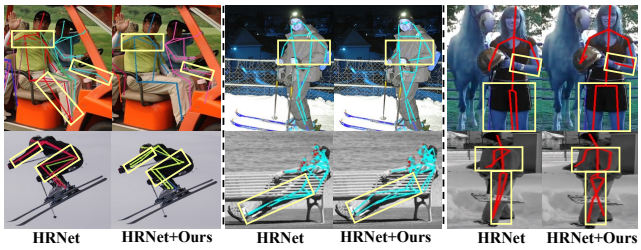


Figure 4: Visual comparisons across three datasets.

Semi-supervised Setting

Quantitative Comparison. We validate ICON in semi-supervised set up in Tab. 4. *'Sup-only'* means supervised model only trained with labeled data. On both cases, the models with ICON outperform *Sup-only* ones at all ratios and exhibit consistent improvement compared with the Pseudo(Lee et al. 2013) and SSPose(Xie et al. 2021a) baselines, which shows that the inter-image consistency supervision can complement the pure location supervision with the learned homogeneous keypoint patterns across images.

Also, ICON has good generality across different ratios of unlabeled data. We observe more accuracy improvements as the labeled data increases. Given different ratios of labeled data, our method gains higher result than SSPose on COCO and MPII. While with 30% labeled data, we can even achieve comparable result $(69.4\%, 85.1\%)$ with the fully-supervised ones $(70.4\%, 88.5\%)$. This reveals that ICON is effective despite in data efficiency.

Dataset	Method	5%	10%	20%	30%	ALL
COCO *(mAP)*	Sup-only	49.2	54.2	58.3	62.7	**70.4**
	Pseudo(Lee et al. 2013)	53.2	57.3	62.2	66.8	-
	SSPose(Xie et al. 2021a)	57.5	61.6	64.5	68.2	-
	Pseudo+ICON	54.0	58.3	63.4	68.3	-
	SSPose+ICON	58.5	62.7	65.6	69.4	-
MPII *(PCKh)*	Sup-only	63.4	69.2	75.1	78.7	**88.5**
	Pseudo(Lee et al. 2013)	67.5	73.5	78.2	81.3	
	SSPose(Xie et al. 2021a)	71.3	76.2	80.3	83.6	-
	Pseudo+ICON	68.4	74.5	79.3	82.5	-
	SSPose+ICON	72.2	77.3	81.5	85.1	-

Table 4: Main results on COCO *val2017* and MPII *val* set.

Ablation Study

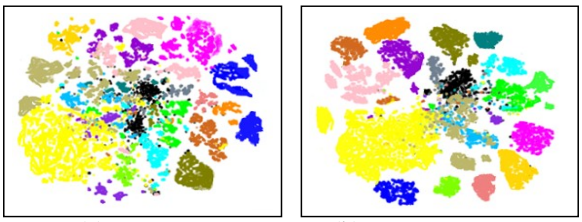
Inter-Image vs. Intra-Image Con. For exploring the effectiveness of the ICON, we develop ablated study of "Intra-Image Con." in Tab. 5, which only samples keypoint pixels within the current instance. It firstly selects queried categories and regards others as negatives, the self as positive to do as in Eq. 1. Result shows that considering the cross-image consistency (*i.e.*, "Inter-Image Con.") achieves more impressive improvement $(+1.7\%)$, compared with baseline. Although "Intra-Image Con." also boosts the performance over baseline by 0.6%, "Inter-Image Con." is preferable.

Single-level vs. Pair-wise. In Tab. 5, We see that the PRCC improves 1.3%, 1.2% than the baseline and it contributes more than SKCC (0.9%) in both cases. It evidences that **i)** the learnt inter-joint instructive relation via PRCC helps deal with the extreme postures and contributes more to the final performance improvement; and **ii)** both keypoint features and pair-wise relations are informative cues. Comprehensively considering both inter-image consistency constraints is essential for the accurate estimation.

Contrast Strategy	Back	MPII*(PCKh)*	COCO*(mAP)*
Baseline	HR32	90.3	74.4
Intra-image Con.	HR32	90.7	75.0
Inter-image Con.	HR32	**92.0**	**76.1**
Base+SKCC	HR32	91.2	75.3
Base+PRCC	HR32	91.6	75.6

Table 5: Study of the different contrastive mechanisms.

Interpretable Visualization of SKCC. To intuitively analyse the SKCC, we visualize the category-specific features over all classes for every image in COCO *val* in Fig. 5. With only MSE loss (a), the category-wise joint features get mixed together, which severely increases the ambiguity among categories and raises the classification difficulty. Equipped with SKCC in Fig. 5 (b), the learned keypoint embeddings become inter-category distinguished and intra-category compact. Thus, more consistent and discriminative keypoint features are remarkably strengthened.



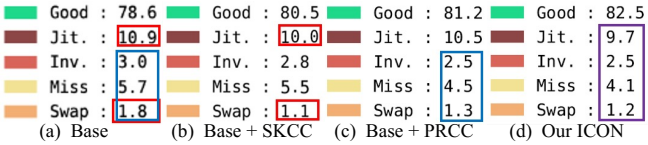
(a) MSE (b) MSE+SKCC

Figure 5: Visualization results of class-wise keypoint embeddings. The different colors denote different categories.

Diagnostic Analysis. For analyzing how ICON contributes to the performance improvement and what problems it exactly solved, we analyze the result with coco analysis tool(Ruggero Ronchi and Perona 2017) in Fig. 6. Four kind of errors are studied:(i) *Jitter* error: small localization error around the correct joint location. (ii) *Miss* error: large localization error, the keypoint locations are not detected completely. (iii) *Inversion* error: confusion between keypoints within an instance. (iv) *Swap* error: confusion between keypoints of different instances.

When used alone, both SKCC and PRCC reduce swap (≥ 0.5) (*e.g.*, swap in Fig. 6 (b), (c) is reduced by 0.7, 0.5). When used together, all types of errors are largely reduced as in (d). ***For SKCC***, the comparison between (a) vs (b) (in red box) shows that making joint features more robust and precise largely alleviates the *jitter* (0.9) and *swap* (0.7) issues. The effectiveness also can be reflected on the visual results in Fig. 7 (b1, b2). ***For PRCC*,** the comparison between (a) vs (c) (in blue box) in Fig. 6 depicts that PRCC

brings significant improvement especially on the large localized errors *i.e.*, *inversion*, *miss* is reduced by 0.5, 1.2. These errors mostly result from the structural occlusions in reality as depicted in Fig. 7, the results in (a1, a2) show that the learned informative relation via PRCC helps infer the missing locations and also guide model correct inverse joints.



(a) Base (b) Base + SKCC (c) Base + PRCC (d) Our ICON

Figure 6: Component analysis in terms of types of errors. The base model adopts ResNet.

Missing Keypoints Reduction. To observe our method for missing error correction, we conduct keypoint-wise evaluation and calculate the keypoint ***Recalls*** at thresholds from 0.1 to 0.9, where each predicted point is viewed as the keypoint if its score is higher than a threshold. The Tab. 6 evidences that overall recall scores of ours are higher than the baseline almost for all thresholds. Particularly, even at the highest threshold (*i.e.*, 0.9), our method outperforms baseline by a large margin, *e.g.*, 5.7% and 6.2% for COCO and CrowdPose. It significantly shows that ICON can decrease the missing errors whether the dateset is crowded or not.

Method	Threshold values								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CO-HR32	99.3	97.5	94.9	91.3	86.9	81.3	72.9	56.2	23.7
+ICON	99.3	**97.7**	**95.3**	**91.9**	**87.9**	**83.1**	**76.1**	**61.0**	**29.4**
CP-HR32	99.3	97.4	94.2	89.8	84.0	76.3	64.7	45.7	13.5
+ICON	99.3	**97.6**	**95.0**	**91.6**	**87.0**	**81.1**	**72.1**	**55.4**	**19.7**

Table 6: Effect of our method for missing error reduction on COCO(CO) *val* and CrowdPose(CP) *val*.

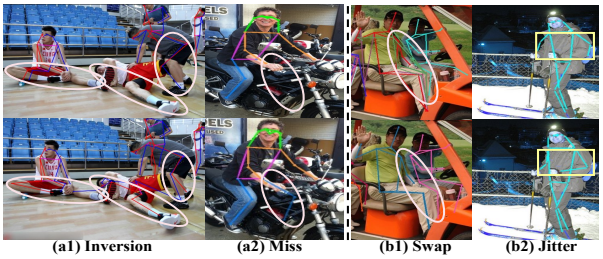

(a1) Inversion (a2) Miss (b1) Swap (b2) Jitter

Figure 7: Result comparisons between baseline (top-row) and baseline with SKCC (*b1, b2*) and PRCC (*a1, a2*).

Sampling Strategy Exploration. We explore the impact of the sampling strategy in Tab. 7. For SKCC, we explore the sampling strategies for the query and negative sampling, only the negative sampling is explored for PRCC because it adopts all pairs as query. We observe that mining more challenging keypoint categories (*i.e.*, "hard"), rather than "random" or "ALL", is indeed useful and more favored to improve the accuracy. It verifies that not all keypoints are necessary and our method is more effective and efficient.

Types	Sampling	Back	COCO	MPII
	Base (w/o ICON)	HR32	74.4	90.3
Single	ALL (17/16)	HR32	74.8	90.7
	Random (5∼10)	HR32	75.0	90.8
	Hard (5∼8)	HR32	**75.3**	**91.2**
Pairs	ALL (8)	HR32	75.1	90.9
	Random (4∼8)	HR32	75.3	91.2
	Hard (4∼5)	HR32	**75.6**	**91.6**

Table 7: Effect of sampling strategy on COCO and MPII.

Numbers of Queries and Negatives. We study the effect of queries and negative sample size (per-class) in Tab. 8. As expected, higher values lead to stronger performance, although from 256 up, the result for query tends to decrease. Since all elements of the selected samples are used during training, the computation and memory complexity increases with a larger size, we selected a size of 256 and 512 for queries and negative samples as a good trade-off.

Types	Back	32	64	128	256	512
Query	HR32	74.6	75.2	75.7	**76.1**	75.9
Negatives	HR32	-	74.5	75.3	75.8	**76.1**

Table 8: Effect of query and negative sample size on COCO.

Discussion. The representation projection head is not used in inference so ICON can be easily applied in other dense prediction tasks to bring substantially improved representation but with no extra overheads.

Besides, the meaning and design of SKCC and PRCC are different. **I)** From a mathematical perspective, if we assume $f_{k_a} = f_{k_b}, f_{k_c} = f_{k_d}$, we can deduce $f_{k_a} + f_{k_c} = f_{k_b} + f_{k_d}$. But if $f_{k_a} \neq f_{k_b}, f_{k_c} \neq f_{k_d}$, we can't deduce $f_{k_a} + f_{k_c} \neq f_{k_b} + f_{k_d}$. Thus, the SKCC and PRCC are not equal actually. **II)** The diagnostic analysis and qualitative results in Fig. 7 (a1, a2) also prove that the semantic relations learnt by PRCC empower the model deal with the structural localization errors than SKCC.

Conclusion

In this paper, we propose a novel framework named inter-image contrastive consistency (ICON), making the first attempt to tackle the inconsistent representations resulting from the large appearance variances across images or occlusions in MPPE. We propose a single keypoint contrastive consistency to semantically drive the feature consistency between keypoints across images, which effectively alleviates location issues. For further dealing with the complex postures (*i.e.*, occlusions), we propose to mine the consistency of keypoint pairs across images for capturing the instructive relation semantics. The ICON can take effect in both supervised and semi-supervised setting without extra overhead for inference on three benchmarks.

Acknowledgements

This work is supported by Beijing Natural Science Foundation (L221012), National Natural Science Foundation of China (61906013, 62106017).

References

Andriluka, M.; Pishchulin, L.; Gehler, P. V.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3686–3693.

Bin, Y.; Cao, X.; Chen, X.; Ge, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Gao, C.; and Sang, N. 2020. Adversarial semantic data augmentation for human pose estimation. *European Conference on Computer Vision*, 606–622.

Bulat, A.; Kossaifi, J.; Tzimiropoulos, G.; and Pantic, M. 2020. Toward fast and accurate human pose estimation via soft-gated skip connections. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 8–15.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.; and Sheikh, Y. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Chaitanya, K.; Erdil, E.; Karani, N.; and Konukoglu, E. 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded Pyramid Network for Multi-person Pose Estimation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7103–7112.

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5386–5395.

Chou, C.; Chien, J.; and Chen, H. 2018. Self Adversarial Training for Human Pose Estimation. *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 17–30.

Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A. L.; and Wang, X. 2017. Multi-context Attention for Human Pose Estimation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5669–5678.

Fang, H.; Xie, S.; Tai, Y.; and Lu, C. 2017. RMPE: Regional Multi-person Pose Estimation. *In Proceedings of the IEEE International Conference on Computer Vision*, 2353–2362.

Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; and Wang, J. 2021. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14676–14686.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. *In Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, J.; Zhu, Z.; Guo, F.; and Huang, G. 2020. The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5699–5708.

Jiang, L.; Shi, S.; Tian, Z.; Lai, X.; Liu, S.; Fu, C.-W.; and Jia, J. 2021. Guided Point Contrastive Learning for Semi-supervised Point Cloud Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6423–6432.

Jin, S.; Liu, W.; Xie, E.; Wang, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Differentiable hierarchical graph grouping for multi-person pose estimation. *European Conference on Computer Vision*, 718–734.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33: 21798–21809.

Ke, L.; Chang, M.; Qi, H.; and Lyu, S. 2018. Multi-Scale Structure-Aware Network for Human Pose Estimation. *In European Conference on Computer Vision. Springer*, 731–746.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems*, 33.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on challenges in representation learning, ICML*, 896.

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; and Lu, C. 2019. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10863–10872.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. *In European Conference on Computer Vision. Springer,*, 740–755.

Liu, S.; Zhi, S.; Johns, E.; and Davison, A. J. 2021. Bootstrapping Semantic Segmentation with Regional Contrast. *International Conference on Learning Representations*.

Luo, Z.; Wang, Z.; Huang, Y.; Wang, L.; Tan, T.; and Zhou, E. 2021. Rethinking the Heatmap Regression for Bottom-up Human Pose Estimation. *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 13264–13273.

Moskvyak, O.; Maire, F.; Dayoub, F.; and Baktashmotlagh, M. 2021. Semi-supervised keypoint localization. *arXiv preprint arXiv:2101.07988*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked Hourglass Networks for Human Pose Estimation. *In European Conference on Computer Vision. Springer,*, 483–499.

Ning, G.; Zhang, Z.; and He, Z. 2018. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *IEEE Transactions on Multimedia*, 1246–1259.

Qiu, L.; Zhang, X.; Li, Y.; Li, G.; Wu, X.; Xiong, Z.; Han, X.; and Cui, S. 2020a. Peeking into occluded joints: A novel framework for crowd pose estimation. *European Conference on Computer Vision*, 488–504.

Qiu, Z.; Qiu, K.; Fu, J.; and Fu, D. 2020b. Dgcn: Dynamic graph convolutional network for efficient multi-person pose estimation. volume 34, 11924–11931.

Radosavovic, I.; Dollár, P.; Girshick, R.; Gkioxari, G.; and He, K. 2018. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4119–4128.

Ruggero Ronchi, M.; and Perona, P. 2017. Benchmarking and error diagnosis in multi-instance pose estimation. *Proceedings of the IEEE international conference on computer vision*, 369–378.

Shi, D.; Wei, X.; Li, L.; Ren, Y.; and Tan, W. 2022. End-to-End Multi-Person Pose Estimation With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11069–11078.

Su, K.; Yu, D.; Xu, Z.; Geng, X.; and Wang, C. 2019a. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5674–5682.

Su, Z.; Ye, M.; Zhang, G.; Dai, L.; and Sheng, J. 2019b. Cascade feature aggregation for human pose estimation. *arXiv preprint arXiv:1902.07837*.

Sun, K.; Lan, C.; Xing, J.; Zeng, W.; Liu, D.; and Wang, J. 2017. Human Pose Estimation Using Global and Local Normalization. *In Proceedings of the IEEE International Conference on Computer Vision*, 5600–5608.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 5693–5703.

Tang, W.; Yu, P.; and Wu, Y. 2018. Deeply Learned Compositional Models for Human Pose Estimation. *In European Conference on Computer Vision. Springer*, 197–214.

Tang, Z.; Peng, X.; Geng, S.; Wu, L.; Zhang, S.; and Metaxas, D. N. 2018. Quantized Densely Connected U-Nets for Efficient Landmark Localization. *In European Conference on Computer Vision. Springer*, 348–364.

Wang, J.; Long, X.; Gao, Y.; Ding, E.; and Wen, S. 2020. Graph-pcnn: Two stage human pose estimation with graph pose refinement. *In European Conference on Computer Vision. Springer*, 492–508.

Wei, S.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional Pose Machines. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4724–4732.

Xie, R.; Wang, C.; Zeng, W.; and Wang, Y. 2021a. An Empirical Study of the Collapsing Problem in Semi-Supervised 2D Human Pose Estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11240–11249.

Xie, Z.; Lin, Y.; Zhang, Z.; Cao, Y.; Lin, S.; and Hu, H. 2021b. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16684–16693.

Yang, W.; Li, S.; Ouyang, W.; Li, H.; and Wang, X. 2017. Learning Feature Pyramids for Human Pose Estimation. *In Proceedings of the IEEE International Conference on Computer Vision*, 1290–1299.

Zhang, J.; Zhu, Z.; Lu, J.; Huang, J.; Huang, G.; and Zhou, J. 2021. SIMPLE: SIngle-network with Mimicking and Point Learning for Bottom-up Human Pose Estimation. *In Proceedings of the AAAI Conference on Artificial Intelligence*, 3342–3350.