DetPoseNet: Improving Multi-Person Pose Estimation via Coarse-Pose Filtering

Lipeng Ke, *Student Member, IEEE* Ming-Ching Chang, *Senior Member, IEEE*
Honggang Qi, *Member, IEEE* Siwei Lyu, *Fellow, IEEE*

***Abstract*—Human detection and pose estimation are essential for understanding human activities in images and videos. Mainstream multi-human pose estimation methods take a top-down approach, where human detection is first performed, then each detected person bounding box is fed into a pose estimation network. This top-down approach suffers from the early commitment of initial detections in crowded scenes and other cases with ambiguities or occlusions, leading to pose estimation failures. We propose the DetPoseNet, an end-to-end multi-human detection and pose estimation framework in a unified three-stage network. Our method consists of a coarse-pose proposal extraction sub-net, a coarse-pose based proposal filtering module, and a multi-scale pose refinement sub-net. The coarse-pose proposal sub-net extracts whole-body bounding boxes and body keypoint proposals in a single shot. The coarse-pose filtering step based on the person and keypoint proposals can effectively rule out unlikely detections, thus improving subsequent processing. The pose refinement sub-net performs cascaded pose estimation on each refined proposal region. Multi-scale supervision and multi-scale regression are used in the pose refinement sub-net to simultaneously strengthen context feature learning. Structure-aware loss and keypoint masking are applied to further improve the pose refinement robustness. Our framework is flexible to accept most existing top-down pose estimators as the role of the pose refinement sub-net in our approach. Experiments on COCO and OCHuman datasets demonstrate the effectiveness of the proposed framework. The proposed method is computationally efficient (5-6x speedup) in estimating multi-person poses with refined bounding boxes in sub-seconds.**

***Index Terms*—human detection, human pose estimation, DetPoseNet, coarse-pose filtering, top-down, bi-directional refinement, unified network, multi-scale learning, multi-stage joint learning, structure-aware loss, keypoint masking, COCO.**

I. Introduction

HUMAN detection and pose estimation from images or videos are two related tasks with many applications in computer vision. While human detection localizes people in bounding boxes, pose estimation further identifies the body parts and skeletal joints of each detected person. The two tasks are intrinsically related to each other. While human detection reduces the search space for pose estimation, pose estimation can also assist human detection in providing constraints on body physique and posture structures that can resolve ambiguities in the scenes. Such bootstrapping is particularly effective in the cases when multiple people occlude one another.

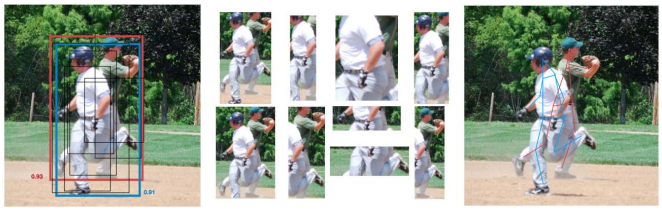

Fig. 1: Difficulties in top-down person detection and pose estimation, where a large number of bounding boxes including mostly redundant ones (in black color) co-exist around valid proposals (red and blue). IoU based non-max-suppression (NMS) in this case tends to keep only *one* (*e.g.* the red) proposal box after NMS, as the blue and red boxes (each for a person under heavy occlusion) are with high overlap. To correctly extract the poses of both individuals, existing top-down methods rely on lowering the NMS threshold to retain more-than-sufficient detection boxes to avoid possible miss-detection. Costly pose estimation must carry out on these redundant proposals which results in high computational cost.

Existing works and systems, however, treat these two tasks separately and usually perform them sequentially. Most approaches [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] choose a **top-down** strategy, where a deep neural network (DNN) model detects all humans, and then another DNN model further estimates the pose of each person. Fig. 1 depicts this top-down approach, where human detection is first performed, and then the detection boxes are fed to a non-max-suppression (NMS) step before they are fed to a pose estimator network one-by-one. Although such top-down methods are simple to implement, they suffer from an *early commitment* problem in human detection — If human detection fails, pose estimation cannot recover the detection errors, and thus the subsequent pipeline of pose estimation will be seriously affected. This is particularly problematic in crowded scenes with heavy occlusions, where the NMS mechanism in human detection tends to remove valid detections due to ambiguities. Specifically in Fig. 1, NMS may merge nearby detection boxes into a single proposal, which is impossible for the pose estimator to recover the miss detection and thus results in poor pose estimation results. Existing top-down methods also tend to suffer from high computational costs in keeping as many detections as possible for pose estimation. As a result, redundant and erroneous pose candidates must be discarded and removed eventually, resulting in a waste of computation.

Lipeng Ke and Siwei Lyu are with the Department of Computer Science and Engineering, University at Buffalo, SUNY, NY, USA, 14260.

Ming-Ching Chang is with the Computer Science Department, University at Albany, SUNY, NY, USA, 12225.

Honggang Qi is with the School of Computer and Control Engineering, University of the Chinese Academy of Sciences, Beijing, China, 101400.

Ming-Ching Chang mchang2@albany.edu is the corresponding author.

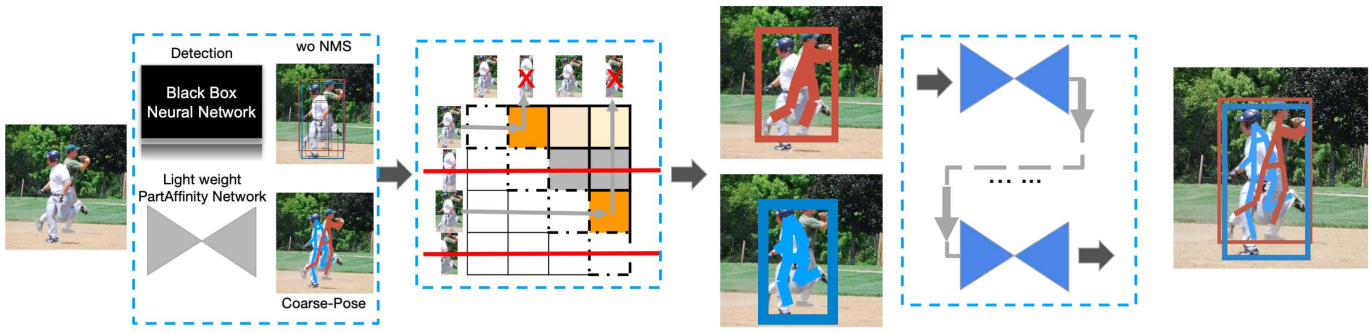

Fig. 2: The **DetPoseNet** multi-person detection and pose estimation is based on a **coarse-pose filtering** design, which consists of three stages of processing: (a) A light-weight *coarse-pose proposal extraction sub-net* consists of two network branches: The *person detection branch* generates human bounding box proposals in a top-down manner. The lightweight *coarse-pose extraction branch* generates coarse human pose proposals using the part-affinity loss in a bottom-up manner. (b) The *coarse-pose based proposal filtering* refines both human detection boxes and pose estimation proposals, which can effectively reduce false-positives and false-negatives. (c) A *multi-scale pose refinement sub-net* of hourglasses stacks takes person detection boxes from the previous stage as input and produces finalized pose skeleton outputs.

In other words, these top-down approaches trade solutions for early commitments with heavy computational costs.

Unlike the top-down approaches, **bottom-up** methods [11], [12], [13], [14] first localize all individual human body keypoints in a single pass and then associate these keypoints to human instances. These methods are robust against early commitments (as they do not rely on human detection). However, bottom-up methods cannot distinguish individual human instances in low pixel resolution, as the network must take the whole image as input to perform keypoint search, and inevitably miss-detections can seriously degrade performance.

In this paper, we present the **DetPoseNet**, a unified, end-to-end framework for human detection and pose estimation. DetPoseNet combines the advantages of top-down and bottom-up methods by leveraging a *coarse-pose filtering* design to improve the performance of both tasks. DetPoseNet consists of three components as in Fig. 2:

1) A light-weight *coarse-pose proposal extraction sub-net* (in Fig. 2a) extracts both the human-bounding-box proposals and coarse-pose proposals in a bottom-up manner without NMS.
2) A *coarse-pose based proposal filtering* (in Fig. 2b) jointly improves the human detection and body keypoint estimations. This coarse-pose filtering replaces the NMS of human detections that are widely used in other top-down methods. Our method basically improves such NMS filtering.
3) A *multi-scale pose refinement sub-net* (in Fig. 2c) finalizes keypoint detections by taking the refined human detection bounding boxes as input.

In our approach, after the initial human detection and coarse-pose proposals are obtained, coarse-pose filtering is applied over all person detection boxes (without NMS), which leads to refined human detection boxes. Pose estimation can then be carried out based on these refined boxes, as in Fig. 1. We will show comprehensive evaluation results of our method and comparison against state-of-the-art methods, including both the top-down human detection and bottom-up pose estimation methods in § IV. DetPoseNet outperforms mainstream methods in multi-human pose estimation on both COCO [15] and OCHuman [16] datasets. DetPoseNet is an extension of our earlier work [17], which only focused on single-person pose estimation. To the best of our knowledge, DetPoseNet is the first end-to-end trainable method that can compete with leading methods on both tasks of human detection and multi-people pose estimation in popular benchmarks.

Our work brings four main contributions:

1) The proposed three-stage DetPoseNet effectively combines the human detection and pose estimation modules into a unified optimization framework.
2) The coarse-pose based proposal filtering module can effectively prevent the early commitment of detection and jointly improve multi-people pose estimation. This design is superior to the bottom-up methods in identifying humans from low pixel resolutions.
3) Pose estimation performance is enhanced with the design of multi-scale supervision and multi-scale regression in the pose refinement sub-net. These designs can simultaneously strengthen context feature learning via a global optimization across scales.
4) Two additional designs of the structure-aware loss and keypoint masking can effectively improve the learning of human body structure in the pose estimation sub-net.
5) The proposed framework can be adapted to any State-of-the-Art top-down pose estimation method, and it can reduce the systematical complexity in terms of pose estimation by 80%.

The remaining of this paper is organized as follows. § II reviews related works and compare our approach to existing methods. § III describes the DetPoseNet in terms of the pipeline and network design details. § IV reports evaluations with results and discussions.

II. Related Works

Human detection and pose estimation have been studied extensively for the past decades. We organize relevant works into three categories: (1) human detection methods that detect and localize people in bounding boxes (§ II-A), (2) single-person pose estimation methods that predict body keypoints based on images cropped from the person boxes (§ II-B), (3) multi-people pose estimation methods that localize each person in the image and further recognize their body keypoints (§ II-C). § II-D compares DetPoseNet to relevant methods in the literature.

A. Human Detection

Modern **human detection** methods are developed upon generic object detector networks that have advanced significantly based on convolutional neural networks (CNNs). R-CNN [18] is a two-stage detector consisting of a proposal generator and an RoI classifier. To reduce the redundant computation of CNN feature extraction from images, region-based feature extraction is introduced in SPP-Net [19] and Fast-RCNN [20]. These methods significantly boost the training and testing speed. The Region Proposal Network (RPN) in Faster-RCNN [21] further speeds up person detection, where the technique is also used in other popular networks such as Mask-RCNN. The Feature Pyramid Network (FPN) [22] generates object proposals at multi-scale layers to resolve the scale mismatches between the RPN receptive fields and the actual object size.

B. Single-Person Pose Estimation

Traditionally, human poses are estimated from images based on local observations of body parts (keypoints) and the spatial dependencies among them. The spatial relationship of articulated human poses is modeled using a tree graph followed by a kinematic chain [23], [24], [25], [26], [27], [1], [28]. Non-tree models [29], [30], [31], [32], [33] augment the tree structure with additional edges to capture relations including occlusion, symmetry, and long-range relationships. To obtain reliable local observations of body parts, CNNs have been widely used to significantly boost human pose estimation [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [17], [47], [9], [8], [10]. Tompson *et al.* [38] use a CNN with a graphical model where parameters are learned jointly by the network. Pfister *et al.* [48] further use CNNs to implicitly capture global spatial dependencies by using networks of larger receptive fields.

The convolutional pose machines (CPM) of Wei *et al.* [35] is a multi-stage architecture based on a sequential prediction framework [49]. It iteratively incorporates global context to refine part confidence maps that can preserve multi-modal uncertainty across iterations. Intermediate supervisions are enforced at the end of each stage to address the problem of vanishing gradients [50], [51], [52] during training. Newell *et al.* [34] also leverage intermediate supervisions that are beneficial to the stacked hourglass architecture. Yang *et al.* [44] design a Pyramid Residual Module (PRM) to enhance the invariance of CNN across scales by learning convolutional filters on various feature scales. However, all these methods make important assumptions that a single person (of interest) is to be detected in the image, and the person should not deviate too much from the assumed location and scale.

C. Multi-People Pose Estimation

The inference of multi-people pose estimation is challenging due to several reasons: (1) unknown number of people in the scene, where each person can appear in various scales and positions, (2) activity or interaction between people can induce complex postures, where the body part articulations and occlusions make the problem difficult. Methods in this category can be organized into *top-down* (§ II-C1) and *bottom-up* (§ II-C2) approaches.

1) Top-down multi-people pose estimation: Top-Down approaches [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] are based on the strategy to first detect each person and then estimate the pose of each person independently in the detected image patch. One major advantage of such approach is that techniques for single-person pose estimation can be directly applicable. However, this simple extension suffers from *early commitments* of the person detection decisions. Methods here also fail to capture spatial configurations among people, which are useful in providing contextual information for leveraging social interaction cues. Lastly, the computation time of these methods grows linearly with the number of people found in the image, as each person's body keypoints must be determined individually via passing through a round of pose estimation network.

2) Bottom-up multi-people pose estimation: Bottom-up approaches [53], [13], [12], [14] employ the strategy to first detect all body keypoints in a single forward of network, and then group these keypoints to human instances. Such *detection-and-grouping* strategy is computationally effective compared to top-down methods, as the computational complexity does not depend on the number of people in the image. [1] Also, the bottom-up approaches are potentially capable of inferring the latent relations among keypoints, which can address the early commitment of person detection issues of the top-down methods. Nonetheless, how to effectively leverage global contextual cues to accurately localize multiple people (who can appear arbitrarily in the view) is still an open research problem. Finally, bottom-up approaches, in general, suffer from low image resolution that can lead to miss-detection of small appearing individuals, as the individuals are too small for the network to detect when the whole input image is resized to fixed input size.

The early work of Pishchulin *et al.* [54] use inter-linear programming to perform greedy association of the fully-connected keypoint candidates. The inference runs slow and can take hours to process a single image. Insafudinov *et al.* [55] improve Pishchulin's method with a more robust

[1]Note that there is a maximum number of people the bottom-up pose estimation method can handle in a single network pass (*e.g.* around 40 people). These methods can deal with the detection and pose estimation of a crowd reasonably well.

backbone using image-dependent pairwise scores to associate the keypoint connections, which significantly reduces the processing time. The Deepercut [55] performs regression on the features extracted from the offset vectors between pairs of body parts. A separate logistic regression is used to convert the pairwise features into probability scores. Cao *et al.* [11] further improve the above pairwise representations using a part affinity field that can learn to associate body parts into the individual human instance. Papandreou *et al.* [12] detect individual body keypoints and predict their relative displacements. A greedy decoding process is used to group the found keypoints into human instances.

D. Comparison to Relevant Methods

In our earlier work of [17], we developed a single-person pose estimator consisting of a set of multi-scale feature supervision and regression modules that learns multi-scale features. Structure-aware loss is used to learn the skeleton configuration in multi-people scenarios and occlusion cases. In § III-C, we will summarize related designs that are integrated into DetPoseNet in the *pose refinement sub-net*.

Regarding our structure-aware loss design that will be described in § III-C3, in contrast to [54] and [55], our approach can efficiently obtain pairwise and triplet occurrences of keypoints without sophisticated design in the training. This capability is sufficient for modeling keypoint relationship for multi-people estimation. A related but independent work is Insafutdinov *et al.* [56] with a simplified body-part relationship graph for faster inference. Their single-frame model is formulated as articulated human tracking via a spatio-temporal grouping of part proposals. Newell *et al.* [13] propose associative embeddings which can be thought as tags representing the grouping of keypoints. Body keypoints are associated via tags with similarities and thus grouped into individual person instances. The Pose Residual Network of Kocabas *et al.* [14] takes person detections and body keypoints as input for the assignment of keypoints to person bounding boxes. Nie *et al.* [57] partition all keypoint detections using dense regressions mapping from keypoint candidates to the person centroids in the image.

We note that Mask-RCNN [7] also generates person detection proposals and performs pose estimation in each proposal region using a single network. We describe two major differences between Mask-RCNN and DetPoseNet: (1) The optimization of Mask-RCNN is *single-directional* in that human box proposals are used to aid pose estimation. In other words, Mask-RCNN can not leverage pose (body parts) to assist the person detection in return. (2) Mask-RCNN is mainly designed for object detection. Thus the pose estimation for each individual person was not improved (but it can be improved as in the DetPoseNet). Specifically, the effective resolution of the proposal of each person in Mask-RCNN tends to be insufficient for body keypoint localization.

III. Method

The proposed DetPoseNet takes an RGB image of size $w \times h$ as input to detect each person, localize body keypoints, and recover the skeletal structure of each person by organizing and linking together body keypoints. The overall pipeline consists of three processing stages (as in Fig. 2): *(1)* In the first stage of *coarse-pose proposal extraction sub-net* (§ III-A), a Faster-RCNN embedded hourglass network generates the person bounding box proposals together with a set of coarse pose proposals; these two kinds of proposals are generated in parallel. *(2)* In the second stage of *coarse-pose based proposal filtering*, the coarse-poses containing rich part layout (of each person instance) are used to refine both the person detection and coarse-pose proposals. We also check if the keypoint proposals are sufficient to form a human skeleton, as incomplete body parts may suggest a false person detection. *(3)* In the third stage, a *multi-scale pose refinement sub-net* (§ III-C) consisting of a light cascaded pose estimator network refines all pose proposals and produces the final skeletal pose estimations.

As aforementioned, top-down methods following the common strategy of first performing person detection and then estimating the pose of each detected person can suffer from early commitment issues. To this end, we propose a *coarse-pose based proposal filtering* in DetPoseNet to address these issues. We will show in § IV that the optimization based on coarse-pose filtering can indeed improve the pose estimation accuracy and performance. Such coarse-pose filtering is superior to the traditional *non-maximal suppression* (NMS) in terms of lowering both the false-negatives and false-positives.

A. Light-weight Coarse-Pose Proposal Extraction

We modify the standard hourglass network [34] into a two-branch *proposal extraction* sub-net that generates both *(i) person detection proposals* and *(ii) coarse-pose estimation proposals* in a single forward pass. As shown in Fig. 2a, the person detection branch extracts person bounding boxes similar to a standard detection network such as Faster-RCNN [21]. The pose estimation branch extracts body keypoint proposals as a set of heatmaps.

The **person detection proposal branch** can be either A black-box or white-box detection network. The former takes the multi-scale features shared from the hourglass network to regress the person bounding boxes on the input image. Compared with the original Faster-RCNN [21], the difference is that we extract features from both Conv-downsampling and Deconv-upsampling layers. The intuition is that features from Conv-downsampling layers contain more contextual information of the input image, while the Deconv-upsampling layers mainly contain high-level features with more information regarding the skeletal structure. We fuse these two sources of features to regress/predict the person bounding boxes. Since this adapted Faster-RCNN considers both individual and inter-body keypoint features, our proposed person detector works jointly with the pose estimator branch by sharing part of the network modules.

The **coarse-pose estimation proposal branch** is an asymmetric hourglass network, which shares a heavy head network with the person detector branch. A set of light tail networks are used to generate keypoint heatmaps (one heatmap per

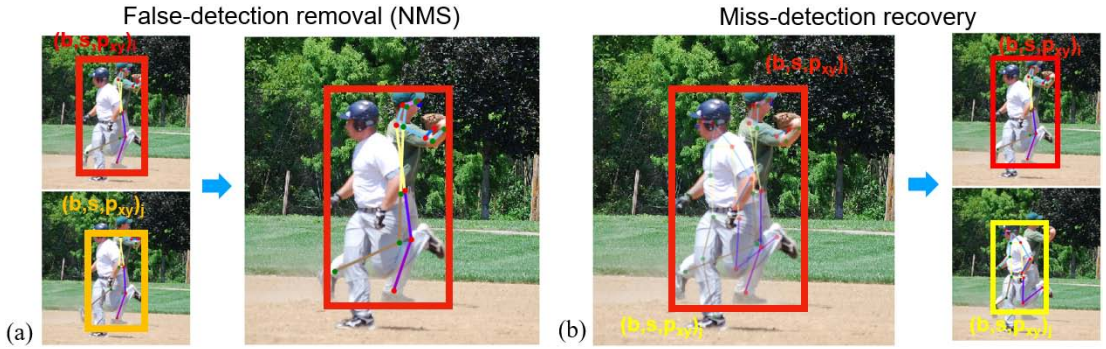

Fig. 3: **Coarse-pose based proposal filtering** refines both human detection proposals and body keypoint proposals in a (box, box-score, pose) triplets, or (b, s, p). (a) shows an example of false-positive removal, where the orange person box is removed. The effect is similar to the standard non-max-suppression (NMS) in existing object detection approaches. (b) shows an example of miss-detection recovery, where the two person instances are identified and are separated into two person instances with identified body keypoints.

keypoint). In this branch, we use the *part-affinity-field loss* [53] to learn the body keypoint associations for each person.

We train the two sub-nets with both the person bounding box and body keypoint annotations as input. The *human proposal loss* term L_p consisting of a human detection loss (L_{det}) and a body keypoint loss (L_{kpt}) in weighted combination are jointly minimized. Specifically, L_{det} is calculated from the loss terms as in Faster-RCNN [21], and L_{kpt} is calculated via the body keypoint alignment loss similar to the part-affinity-field loss [53].

B. Coarse-Pose Based Proposal Filtering

The optimization strategy of state-of-the-art multi-person pose estimation works is either top-down or bottom-up. Top-down optimization *e.g.* Mask-RCNN [7] estimates human pose according to detection proposals and tends to suffer from false-detections. On the other hand, bottom-up optimization detects and associates body keypoints and tends to suffer from miss-detections due to low image resolution. To the best of our knowledge, this work is the first to incorporate both top-down and bottom-up optimizations into a unified framework. We proposed a **coarse-pose based** proposal filtering strategy that combines advantages of both top-down and bottom-up optimization schemes to better address the false-detection and miss-detection issues.

Specifically, our coarse-pose based proposal filtering aims to: (1) eliminate redundant human detection box proposals using body keypoint proposals from the pose estimation branch, and (2) recover possibly missing (false-negative) person detections by predicting a putative bounding box from the detected body keypoints. The former is essentially the improvement of human detection using pose estimations, and the latter is the improvement of pose estimation using human detection. This way, the coarse-pose based proposal filtering combines the advantages of both *bottom-up* pose estimation and *top-down* human detection, such that latent relationships between human bounding boxes and body keypoints can be exploited and leveraged together in a single optimization framework. This way, false positive detections can be effectively reduced, and false negative detections can be recovered. Fig. 3 illustrates an example of this two-way optimization. In Fig. 3a, false positive detections are eliminated if there is not sufficient keypoints to form a skeleton, or if the pose estimation is similar to any identified pose instance. In Fig. 3b, false negative detections can be recovered by creating a new bounding box from the remaining identified body keypoints.

Algorithm 1: Coarse-pose based proposal filtering

Input: person detection boxes $\mathcal{B}$, keypoint heatmaps $\mathcal{K}$
Output: refined person box & pose set $\tilde{\mathcal{P}}$ initialized as $\emptyset$

Initialisation : Sort $\mathcal{B}$ by detection scores decreasingly, initialize $\tilde{\mathcal{P}} = \emptyset$.
$\mathcal{P} = PoseExtractor(\mathcal{B}, \mathcal{K})$ // sorted set of (b, p)
for each $(b, p) \in \mathcal{P}$ **do**
  if $n_{kpt}(p) \leq \delta_{kpt}$ **then**
    continue //likely false detection
  end if
  if $n_{kpt}(p) > \delta_{kpt}$ and $\exists (b_j, p_j) \in \tilde{\mathcal{P}}$ s.t. $b = b_j$ **then**
    //recover a likely miss detection
    $b' = kpt2box(p)$ //new box
    $\tilde{\mathcal{P}}$.append (b', p)
  else
    //normal case: b matches observed keypoints
    $\tilde{\mathcal{P}}$.append (b, p)
  end if
end for
return $\tilde{\mathcal{P}}$

The *top-down* optimization in Fig. 3a can also be regarded as an improved human pose-based NMS. In existing methods, such NMS is typically performed by checking the *intersection-over-union* (IoU) of detection boxes. This standard approach is required by many other methods to eliminate duplicated and highly overlapping responses around the peak of each detection location. However, simple IoU-based NMS relies on a critical assumption, that only a *single* person exists around the peak of the detection. Such assumption restricts the ability to distinguish or resolve ambiguities of crowded or occluded cases, and thus can lead to potential false negative detections. For the example in Fig. 3, the two baseball players

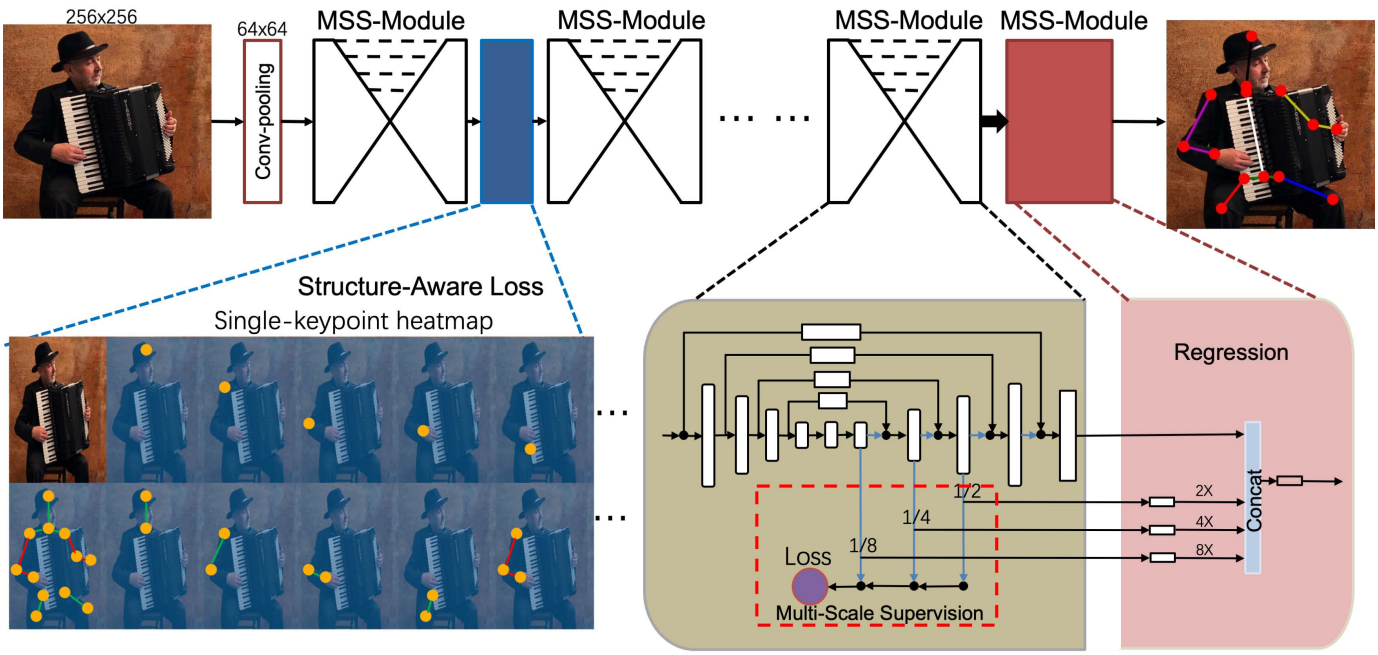

Fig. 4: The multi-scale **pose refinement sub-net** incorporates a structure-aware design that can accurately localize the body keypoints of each person, by taking the person person bounding box after coarse-pose based filtering as input. It consists of three components: *(i)* multi-scale supervision (MSS) module (§ III-C1), *(ii)* multi-scare regression (MSR) module (§ III-C2), and *(iii)* intermediate supervision layers using the *structure-aware loss* (§ III-C3). This network pipeline is fine-tuned using the *keypoint masking training* scheme (§ III-C4).

are visually overlapping. Thus the multiple responses of highly overlapped detection boxes will be blindly removed by typical NMS until one detection is left, which results in false negative detection of the other player. In comparison, our coarse-pose based proposal filtering can effectively leverage identified body keypoints from pose estimation to avoid such error.

Algorithm 1 shows the pseudo code for the proposed coarse-pose filtering. Let $\mathcal{B} = \{b_1, \cdots, b_J\}$ denote the set of person detection box proposals (J is the total number of box proposals), where each box $b_i = (x_i, y_i, w_i, h_i, c_i)$. The human body keypoint heatmap $\mathcal{K}$ is represented as a third-rank tensor $(W \times H \times M^\star)$, where $M^\star$ is the total number of heatmaps. Initially there are $M^\star = M$ heatmaps (M denotes the number of human body keypoints). In the later stages of our algorithm when we consider structure-aware loss, additional heatmap channels across adjacent keypoints will be considered. Let $\tilde{\mathcal{K}}$ denote ground-truth heatmaps.

The person detection box proposals in $\mathcal{B}$ are first sorted by their confidence scores c_i. Then bottom-up pose estimator $PoseExtractor$ extracts the (box, pose) tuple (b, p) from detection proposals $\mathcal{B}$ and the heatmap map $\mathcal{K}$. For each tuple (b, p) in the human box-pose set $\mathcal{P}$, function $n_{kpt}(p)$ checks if pose p contains sufficient number of identified keypoints δ_{kpt} such that p represents a valid human pose skeleton. Next we determine if the pose p already exists in refined pose set $\tilde{P}$, by checking the COCO Object-Keypoint-Similarity (OKS) [15] metric. If p is not in $\hat{BP}$ (set of (b, p)), the corresponding box b will be regarded as false detection. Finally, for each person detection box, we check if there is another pose p_j in $\tilde{P}$ that is with less similarity than a threshold and is located within the same bounding box. If so, both boxes b_j and b_i will be used as input to fed into the pose refinement sub-net to recover pose skeletons.

C. Multi-Scale Pose Refinement Sub-net

The multi-scale **pose refinement sub-net (PRS)** in Fig. 2(c) consists of light-weight cascaded hourglasses in stacks that finalize the body keypoint localization for each person, based on refining keypoints identified from the previous stage. The PRS features two *multi-scale* designs and two *structure-aware* training strategies, in which part of the designs are reported in our previous work of [17] for single-person pose estimation. Here we adopt these for the detection and pose estimation of multiple individuals after the coarse-pose based filtering in the new pipeline. Details are described in the following.

The *multi-scale supervision* (MSS) module (§ III-C1) is an enhancement of the standard hourglass conv-deconv network with skip connections [34] that can be trained with multi-scale loss supervision. The *multi-scale regression* (MSR) module (§ III-C2) performs a pose structural regression by matching multi-scale keypoint heatmaps and their high-order associations, which produces the final human pose estimation at the end. As in Fig. 4(top), MSS stacks can be repeated multiple times depending on the available GPU memory. Both the MSS and MSR modules share a common *structure-aware loss* function (§ III-C3), which is designed to ensure effective multi-scale structural feature learning. The training of the whole pipeline is fine-tuned using the *keypoint masking training* scheme (§ III-C4) to focus on learning hard samples.

We next compare and motivate our multi-scale and structure-aware design, and highlight the differences with other

standard hourglass networks. First, our conv-deconv hourglass stacks aim to capture rich features for keypoint detection across large variability in appearances and scales. In contrast, standard hourglass [34] is sensitive to a particular scale in the multi-scale pyramid, and thus lacks a robust and consistent response across scales. To this end, we explicitly add layer-wise supervisions to each of the deconv layers in the training of our MSS module.

Secondly, our MSS hourglass model can learn and infer refined keypoint heatmaps by considering the global structures. This can be observed by comparing the MSS heatmaps before and after training. Recall that each heatmap corresponds to the location likelihood of each body keypoint (*e.g.*, elbows, wrists, ankles, knees). During the training of our MSS, heatmaps are supervised against the ground-truth body keypoint heatmaps that are typically generated using 2D Gaussian blurring as initialization. After training, at the test runs of the MSS pose estimation, we observed mostly non-Gaussian heatmaps that variate according to the human gestures in the training images. In contrast, a key deficiency in the original hourglass model [34] is that each keypoint heatmap is estimated *independently* in the way the relationship between keypoints are not considered. In other words, structural consistency among detected keypoints is not optimized in these methods.

Thirdly, to ensure structural consistency in the pose estimation pipeline, the *structure-aware loss* supervision is introduced in between the MSS hourglass modules, which serves as the purpose of intermediate supervision. This can better capture the adjacency and associations among the body keypoints. The structure-aware loss is also used in the MSR module at the end of the pipeline, to oversee all keypoint heatmaps across all scales globally. This way a globally consistent pose configuration can be inferred as the final output.

Finally, the MSR matches both individual body keypoints (first-order consistency) and pairwise consistencies among adjacent keypoints (second-order consistency). The co-occurrence of a matching pair between a hand/leg *w.r.t.* the head/torso with high confidence can provide a stronger hypothesis that can win over other separated, uncorrelated individual matches. This way, global structural posture consistency can be better inferred. The MSR module is trained to perform such optimization across all body keypoints, all scales of features, and all pairwise correlations in a joint regression.

1) Multi-scale supervision (MSS): To effectively learn deep features for each scale and across multiple scales, multi-scale supervision (MSS) is performed during model training at all deconv layers. We explicitly match the keypoint heatmaps at each scale with the corresponding down-sampled ground-truth heatmaps, and compute the *residual* at each deconv layer to calculate the loss (*e.g.*, at 1/2, 1/4, 1/8 down-sampling scales). The MSS network architecture is depicted in the gray box at the bottom-right of Fig. 4. To make equal the feature map dimensions to compute the residual at the corresponding scale, a 1-by-1 convolutional kernel is used for dimension reduction. The 1-by-1 conv also converts the high-dimensional deconv feature maps into the desired number of features, where the number of reduced dimensions matches the number of body keypoints (*i.e.* the number of heatmaps).

Fig. 5: **Multi-scale supervision (MSS)** can provide keypoint refinement analogical to the *attention* mechanism in resolution pyramid search. (a) to (c) are the 8x, 4x, 2x scales keypoint heatmaps respectively. The progression from (a) to (c) shows the refinement of keypoint heatmaps during the deconv up-sampling, where the location of the thorax keypoint is refined with increasing accuracy and more concentrated heatmap.

The multi-scale supervision can provide keypoint localization refinement similar to the *attention* mechanism [58] of conventional resolution pyramid for image search. The activation areas in the low resolution heatmap can provide guidance of the location refinement in the subsequent high-res layers. Fig. 5 provides a visual illustration.

We next describe the loss function L_{MSS} for the training of the multi-scale supervision module. L_{MSS} is defined by summing the L_2 loss from the heatmaps of all keypoints across all scales, similar to the multi-scale loss function used in [35], [34]. To detect the $M = 16$ keypoints (head, neck, pelvis, thorax, shoulders, elbows, wrists, knees, ankles, and hips), the M heatmaps are generated after each conv-deconv stack. The loss at the s-th scale is calculated by comparing the predicted heatmaps of all keypoints against the ground-truth heatmaps at each matching scale:

$$L_{MSS} = \frac{1}{S}\sum_{s=1}^{S} L_{SA}(K^{(s)}, \tilde{K}^{(s)}) \tag{1}$$

where $K^{(s)}$ and $\tilde{K}^{(s)}$ represent the ground-truth and the predicted confidence maps at the scale s for all the keypoint, respectively.

In standard datasets [59], [15], ground-truth poses are provided as the keypoint locations. We follow the common practice for generating ground-truth heatmaps as in Tompson *et al.* [37], where the i-th keypoint ground-truth heatmap K_i is generated using a 2D Gaussian centered at the keypoint location (x, y), with a standard deviation of 1 pixel. Fig. 4 (bottom left, first row) shows a few examples of the ground-truth heatmaps for certain keypoints.

2) Multi-scale regression (MSR): A *fully convolutional* multi-scale regression (MSR) is performed after the MSS conv-deconv stacks to refine the multi-scale keypoint heatmaps globally. This can effectively improve the structural consistency of the estimated poses. The intuition is that the relative positions of arms and legs *w.r.t.* the head/torso provide useful action priors, which can be learned from the regression network by considering feature maps across all scales for pose refinement. The MSR module takes the multi-scale heatmaps as input and matches them to the ground-truth keypoints at respective scales. This way the regression network can

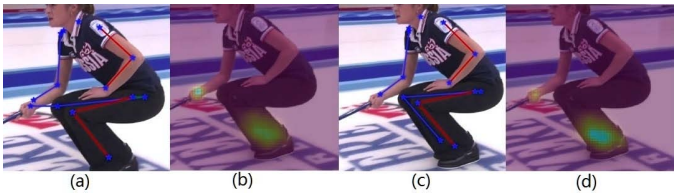

Fig. 6: **Muti-scale regression (MSR)** of body keypoints to disambiguate multiple peaks in the keypoint heatmaps. (a-b) shows an example of (a) keypoint prediction and (b) heatmap from the MSS module hourglass stacks, which will be fed into the MSR module for regression. (c-d) shows (c) the output keypoint locations and (d) heatmap after MSR. Observe that the heatmap peaks in (d) are more focused compared to (b).

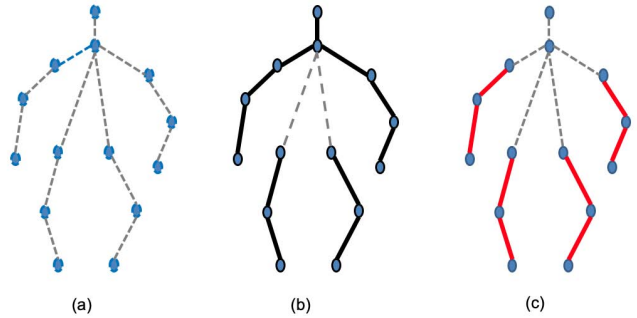

Fig. 7: The **human skeleton graph** $\mathcal{G}$ used for structure-aware loss definition: (a) Blue dots depict body keypoints. (b) Thick black lines depict pair-wise connected keypoints and limbs. (c) Red lines depict the triplets of keypoint connections for calculating the structure-aware loss for elbows and knees.

effectively combine heatmaps across all scales to refine the estimated poses.

The MSR module jointly optimizes the global body structure configuration by determining the connectivity among body keypoints based on mutli-scale features. The effect of MSR can be regarded as an extension to the work of Convolutional Part Heatmap Regression [42], where only keypoint heatmap regression at the original scale of the input image is considered. The input image with the keypoint heatmaps can be viewed as an attention mechanism (as in MSS) commonly used in the resolution pyramid. In this view, our MSR module learns a scale-invariant, attention-based structural model with better performance. Moreover, our MSR module optimizes the structure-aware loss, which matches individual keypoints and higher-order associations (pairs and triplets of keypoints) for pose estimation. The output from the MSR module is thus a comprehensive pose estimator capable of considering pose configurations across multiple feature scales, multiple keypoint associations, and high-order keypoint associations.

Fig. 6 shows an example with improvements brought by the MSR module. MSR works hand-in-hand with the MSS module to explicitly model high-order relationship among body parts, such that posture structural consistency can be maintained and refined.

3) Structure-aware loss: It is a consensus that deeper CNN hourglass stacks lead to better pose estimation results [34]. As the depth of hourglass stacks increases, *gradient vanishing* becomes a critical issue in training the network, where *intermediate supervision* [35], [34], [43], [44] is a common solution. In this regard, we design a structure-aware loss function following the intrinsic human skeletal graph structure. Such structure-aware loss design was implemented in two places of our network: (1) in-between the MSS module stacks as a means of *intermediate supervision* to enforce structural consistency while localizing keypoints; and (2) in the MSR module to determine a globally consistent pose configuration.

The structure-aware loss is calculated according to the *human skeletal graph* $\mathcal{G}$ shown in Fig. 7. Each node $\mathcal{N}(n)$ (blue dots) represent a body keypoint of the human skeleton and its connected keypoints, $n \in \{1, ..., M\}$. Thick black lines depict pair-wise connected keypoints and limbs. Red lines depict the triplet connections for the joints of elbows and knees. The *structure-aware loss* L_{SA} for each scale is calculated as:

$$\frac{1}{M}\sum_{n=1}^{M}\left(||K_{:,:,n}-\tilde{K}_{:,:,n}||_2+\alpha_2\sum_{n'\in\mathcal{N}(n)}||K_{:,:,n'}-\tilde{K}_{:,:,n'}||_2\right), \tag{2}$$

where the first term calculates individual keypoint matching loss, and the second term represents the structural matching loss, in which $K_{:,:,n'}$ and $\tilde{K}_{:,:,n'}$ are the ground-truth and prediction heatmaps for individual keypoint n and its neighbors in graph $\mathcal{N}$, respectively. Hyperparameter α_2 is a weight balancing the two terms.

All structural connectivity of keypoints is empirically determined to match the human skeletal graph $\mathcal{G}$ to better capture the physical connectivity of the human body as structural priors. The structure-aware loss is typically calculated upon pairs of connected keypoints, *e.g.*, head-thorax, shoulder-elbow, wrist-elbow, hip-knee, hip-hip, knee-ankle, in the bottom subfigure of Fig. 4. Since the elbows and knees have multiple physical connections (in contrast to the shoulders and wrists, and the hips and ankles, respectively), and the arms/legs can be easily occluded due to their no-rigid flexibility compared with the torso, we enforce *three-way* structure-aware loss for elbows and knees (*e.g.*, hip-knee-ankle, shoulder-elbow-wrist).

Fig. 4 (bottom left) shows a breakdown visualization of how our skeleton-guided structure-aware loss is calculated in traversing the keypoints and their relationships according to $\mathcal{N}$. The top row in the sub-figure shows the intermediate loss defined on individual keypoints (*e.g.*, the right ankle, knee, hip, pelvis, thorax, head, wrist, elbow), where similar designs are used in [35], [34]. The bottom row shows our structure-aware loss defined for a set of connected keypoints, which is unique in our method.

4) Keypoint masking in training: Occlusion or partly observed body keypoints can strongly affect the performance of human detection and pose estimation, especially for crowded or multi-people scenarios. While the pose estimator is optimizing the search and localization of each body keypoint, there might exist multiple suitable keypoints in the vicinity (due to multiple subjects), or it could appear that none is found due to occlusion. The structure-aware loss from § III-C3 can address some of this challenging issue. Here we further enhance our model capability during training via a *data augmentation* scheme by masking keypoints.

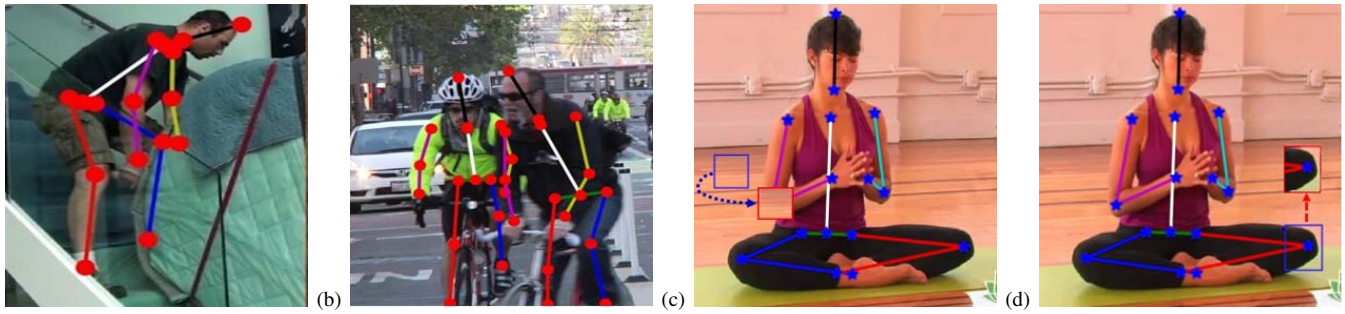

Fig. 8: **Keypoint masking** to simulate hard training samples that can effectively improve robustness in body keypoint localization and association for pose estimation in crowded scenes. (a) is a common case in human pose estimation, where the keypoint (left-wrist) is occluded but can still be localized. (b) is a case where the nearby person's keypoint can be mismatched. We propose two keypoint masking strategies: (c) **background keypoint masking** to address the issue of (a) by cropping a background patch and pasting onto a human body keypoint, and (d) **keypoint duplication masking** to address the issue of (b) by cropping a body keypoint patch and pasting it onto the background. The approach of (c) simulates the cases when body keypoints are invisible, and (d) simulates the multi-people or crowded scenarios with multiple peaks in the keypoint heatmap.

To motivate, refer to Fig. 8a, where the left wrist of the person is completely occluded, however such occlusion can still be recovered from the detected left elbow and arm via structural connectivity. Fig. 8b shows another difficult case where multiple people (and thus multiple body keypoints) appear nearby, thus the pose estimator may mix and connect keypoints across the nearby people and produce erroneous results. From a data-driven point of view, a typical limitation of the currently available human pose dataset is that the amount of annotated samples for such complex scenarios is very limited and insufficient to train a deep network. Conventional data augmentation methods (such as image transformations or color jittering) do not effectively address these issues. Our aim here is to develop an effective data augmentation approach that can address the issues of keypoint occlusions and multiple keypoints.

In our **keypoint masking data augmentation**, we perform copy/paste of selected keypoint patches from the image to simulate two effects: *keypoint occlusion* and *keypoint ambiguity* (due to additional keypoints from other people) in the vicinity. We implemented the following two types of keypoint data augmentation to generate hard training samples to improve training: (1) **background keypoint masking**, by copying a background patch and pasting onto a body keypoint to cover it (as in Fig. 8c) to simulate keypoint occlusion. This augmentation is useful for the learning of occlusion recovery. (2) **keypoint duplication masking**, by copying a body keypoint patch and pasting onto a nearby background (as in Fig. 8d) to simulate the multiple keypoint ambiguity in the vicinity. These augmentations can effectively simulate the frequent ambiguities occurring in multi-people pose estimation. Since these augmentations produce multiple identical keypoint patches, the trained model can better learn to infer keypoint ambiguities for multi-people cases. The keypoint masking training is especially beneficial to fine-tune our keypoint regression network across multiple scales of features. Finally, the proposed keypoint masking data augmentation can be easily performed using known ground-truth keypoint annotations — it can be performed either online (dynamically) or offline (as pre-processing) for network training.

IV. Datasets and Evaluations

We evaluate our method on three public benchmarks for the joint tasks of simultaneously human detection and pose estimation: (1) **COCO keypoint challenge** dataset [15] with annotations of 17 keypoints for each person (12 body parts and 5 facial landmarks). The COCO training set consists of over 100K person instances with over 1 million labeled keypoints. The test set contains "test-challenge" and "test-dev"subsets, where each contains roughly 20K images. (2) **OCHuman** dataset [16], which contains three types of human-related annotations: detection bounding boxes, instance binary masks, and 17 body keypoint locations. OCHuman is particularly challenging and suitable for evaluating our method, as the subjects in the dataset are usually heavily occluded by one or several other people. On average, each person bounding box contains 0.67 IoU overlapping with boxes of nearby people. In comparison, the average IoU for each person is only 0.01 in the COCO dataset. The OCHuman dataset is split into $2,500$ validation and $2,231$ testing images, which contains $4,133$ and $3,797$ human instances, respectively. Images in these datasets are collected from diverse scenarios with real-world activities including crowds, large scale and view variations, occlusions, and interaction among people when performing various movements.

We compare DetPoseNet against popular methods including OpenPose [11] and Mask-RCNN [7] on COCO evaluation in Table I. We report results on COCO keypoint challenge and ablation study in Table II and performance comparison on OCHuman in Table III. Experiments on the above two datasets show that our DetPoseNet significantly outperforms the previous state-of-the-art methods.

For qualitative results, Figures 10 and 11 show visual results of the DetPoseNet on various real-world challenging scenes, including the difficult cases of large scale variations of individuals, large viewpoint variations, heavy body keypoint occlusions, highly overlapping body layouts among interactions or group activities, crowded scenes, large illumination and appearance variations, and diverse activities with distinct body layouts and movements. In many cases, DetPoseNet can

TABLE I: Results on the COCO test-dev dataset for both top-down and bottom-up approaches. AP^{50} denotes OKS $= 0.5$. AP^{L} denotes evaluation for large-scale people only.

Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Bottom-Up Approaches					
METU [14]	70.5	87.7	77.2	66.1	77.3
PersonLab [12]	68.7	89.0	75.4	64.1	75.5
Associative Emb. [13]	65.5	86.8	72.3	60.6	72.6
OpenPose [11]	64.2	86.2	70.3	61.2	68.7
Top-Down Approaches w/ External Detector					
MSPN [47]	**76.1**	**93.4**	**83.8**	**72.3**	**81.5**
MRSA [9]	73.7	91.9	81.1	70.3	80.0
HRNet-W48 [10]	75.5	92.5	83.3	71.9	81.5
EvalPose [60]	75.7	91.9	83.1	72.2	81.5
Top-Down Approaches w/o External Detector					
RMPE [6]	73.3	89.2	79.1	69.0	78.6
G-RMI [5]	71.0	87.9	77.7	69.0	75.2
Mask R-CNN [7]	69.2	90.4	76.0	64.9	76.3
DetPoseNet	**75.3**	**95.2**	**83.1**	**71.5**	**80.9**

TABLE II: Results COCO val dataset.

Method	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
GT Bbox + PRS	78.1	92.6	84.0	74.1	83.3
F-RCNN + PRS	75.0	88.7	80.9	69.4	82.9
F-RCNN w/o FP + PRS	75.3	89.6	81.7	70.6	83.1
F-RCNN w/o FN + PRS	76.1	91.0	82.7	72.1	83.1
DetPoseNet	77.3	92.1	83.8	73.6	83.3
DetPoseNet w/ HRNet-48	**78.3**	**92.7**	**84.2**	**74.3**	**83.5**

† F-RCNN denotes Faster-RCNN [21]. PRS denotes the proposed *pose refinement sub-net*. *F-RCNN w/o FP* denotes Faster-RCNN detected box without false-positive from human detection. *F-RCNN w/o FN* denotes Faster-RCNN without without false-negative from human detection.

detect and localize all individuals with highly accurate body part localization.

A. Results on COCO Keypoints Detection

The COCO keypoint challenge evaluation is performed based on the Object Keypoint Similarity (OKS) metric [15], which is similar to the IoU metric in the evaluation of object detection. The COCO evaluation results are reported using the mean average precision (AP) over 10 OKS thresholds as the main competition metric. The OKS is calculated from the scale of each person and the distance between the predicted and groundtruth points. Table I shows results from top teams in the challenge. DetPoseNet achieves the best scores over all state-of-the-art comparison methods that do not rely on external person detectors.

Table II reports the comparison and ablation study results on the COCO validation set. First of all, an experiment is conducted by feeding the ground-truth person detection boxes to our pose refinement sub-net (PRS), which represents a reduced problem of single-person pose estimation based on our earlier work in [17]. As a result, DetPoseNet achieved the AP of 78.10%. This result can be regarded as the upper-bound of the top-down pose estimation approach without the use of the proposed coarse-pose proposal filtering. In a following up experiment, we use Faster-RCNN [21], one of the state-of-the-art object detectors, to produce person detection boxes, and continue the pose refinement sub-net in a similar setup. As a result, the AP drops down to 75.0%. Finally, in a third experiment as in the last row of Table II, with the adding of the proposed coarse-pose proposal filtering on top of the Faster-RCNN, DetPoseNet obtained AP of 77.3, which is comparable with the use of ground truth person boxes. If we further remove the false positive detection box proposals, our DetPoseNet outperforms the groundtruth box result by 0.6%. Moreover, when we replace the use of HRNet-48 [10] as the pose refinement sub-net in DetPoseNet, the performance can be further improved to 78.3, which shows that our proposed framework can be adapted to the latest single person pose estimator.

Results in Table II also indicate the following observations. (1) performance of the top-down human pose estimation approaches relies heavily on the person detector as input. Imprecise person bounding box induces performance bottleneck. Top-down pose estimation approaches suffer from both performance drops and increased computational complexity due to the false positive detections and false negative detections arising from imprecise input boxes. (2) Our results also show that the use of ground truth person bounding boxes may not represent the performance upper-bound for top-down multi-people pose estimation, as the manually annotated groundtruth boxes may not be the "perfect" input localization for a data-driven learned pose estimation network. To this end, our proposed DetPoseNet is a jointly-learned, end-to-end framework for both human detector and pose estimator that can achieve superior performance. (3) Performance of pose estimator generally increases with its network depth.

TABLE III: Performance comparison on the **OCHuman** dataset.

	OCHuman	Detection			Pose		
		mAP	AP^{50}	AP^{75}	mAP	AP^{50}	AP^{75}
Mask-RCNN	val	0.263	0.502	0.258	0.112	0.248	0.086
DetPoseNet	val	**0.381**	0.662	0.174	**0.380**	0.663	0.260
Mask-RCNN	test	0.260	0.503	0.128	0.109	0.244	0.082
DetPoseNet	test	**0.373**	0.627	0.261	**0.378**	0.631	0.258

B. Results on OCHuman Dataset

To demonstrate how DetPoseNet performs on occluded scenarios, we report person detection and multi-person pose estimation results on the challenging OCHuman dataset in Table. III. DetPoseNet outperforms Mask-RCNN [7] in both detection and pose estimation tasks by a significant margin. This shows that our proposed *coarse-pose proposal filtering* approach can handle crowded scenes and heavy occlusions much better than the typical top-down approach.

C. Ablation study on COCO validation set

We next investigate how the *coarse-pose proposal filtering* can be applied to (other) *generic* human detectors to improve human pose estimation. In the following experiments, we feed the human detection proposals from other mainstream detectors as input to our coarse-pose proposal filtering pipeline, and compare the resulting pose estimation performance gain. Fig. 9 summarizes these results. The x-axis denotes the accuracy of mainstream human detectors (left to right: SSD300, SSD512, retinanet, faster_rcnn_r101, cascade_r50, htc_r50, htx_x101, htc_fpn) ranked order by their detection mAPs. The y-axis

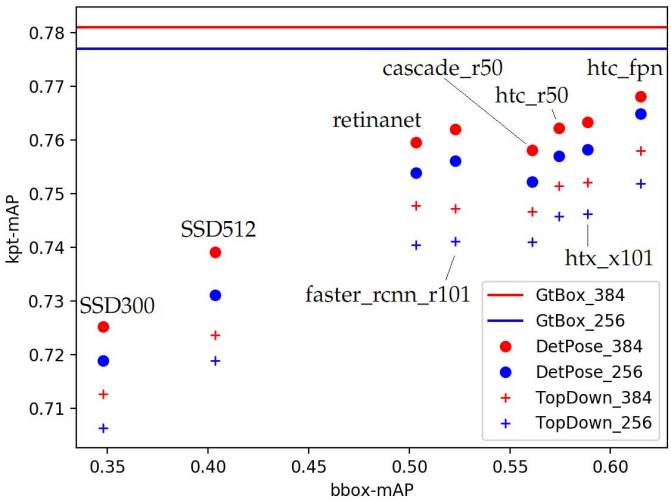

Fig. 9: Evaluation on applying the **coarse-pose based proposal filtering on mainstream human detectors** as input on the COCO validation set (see text). Results show that the proposed coarse-pose proposal filtering can consistently improve pose estimation when combined with any of these top-down human detectors.

denotes the pose estimation mAP produced using the respective person boxes as inputs, with ('·') and without ('+') coarse-pose based proposal filtering, in two resolutions (256x256 and 384x384) in the final computation stage. The two horizontal lines denote the (upper bound) pose estimation accuracy using ground truth human boxes as input.

Fig. 9 shows that for all human detectors that are evaluated, our *coarse-pose proposal filtering* consistently improves the joint pose estimation mAP. Another observation is that, along each curve, the detection mAP does not monotonically increase as the human detection mAP increases. Relationships between person detection boxes (the inputs) and pose estimation responses is intricate but consistent.

D. Efficiency Analysis

We formulate the computation complexity of our **multi-person** pose estimator as $\mathcal{O}(det)+N*\mathcal{O}(pose)$, where $\mathcal{O}(det)$ and $\mathcal{O}(pose)$ are the complexity of the person detector and the **single-person** pose estimator, respectively; and N is the times of single-person pose estimator being run. This way computational efficiency can be compared systematically and fairly. As discussed in Sec. I, top-down approaches usually suffer from the early commitment of person detector; thus, in these approaches, a single-person pose estimator is applied much more than the number of actual people in the images (*i.e.* a much larger N is required). To illustrate, there are $11,004$ people in the COCO validation set, however HRNet [10] and MSRA [10] apply 9.5 times $(104,125)$ of single-person pose estimator runs, while MSPN [47] applies 8.0 times $(88,291)$ of single-person pose estimator runs. In comparison, in our proposed framework, we achieve similar or better performance using only 1.6 times $(17,602)$ of single-person pose estimator runs. Thus, with our proposed framework, the inference in terms of the pose estimation can be sped up 5 to 6 times. Our proposed framework is open and versatile, so it can be adapted to and combined with recent top-down single-person pose estimation methods, including the new ones, to leverage newer advantages and breakthroughs.

E. Failure Case Analysis

Fig. 12 shows the failure cases of DetPoseNet on the COCO evaluations. Occlusions of body parts can lead to localization errors or miss-detections. Such occlusion-related problems can also be addressed back to the data annotation quality, in which we found that these occluded keypoints should be labeled in mainstream datasets but in reality they were not. In highly crowded scenes with overlapping people, the pose estimator tends to confuse or miss keypoints from different people. We found that humanoid statues or animals can be wrongly detected as people in some cases. These issues could be mitigated by adding more negative examples in the training set. We also found that repeating multiple runs with slight changes of scales and rotations can sometimes yield superior results. However such "trick" can also reduce detection accuracy. In our experiments, global accuracy on the COCO validation set can drop by 5% this way. We avoid such approaches based on repeated trials followed by decision fusion, as these tricks are not a principal solution for real-world applications.

V. Conclusion

Human detection and multi-people pose estimation is an important tasks for human-centric visual understanding and AI applications. In this work, we proposed DetPoseNet, a unified human detection and pose estimation pipeline, that can jointly optimize the two tasks in a hybrid pipeline. The DetPoseNet can effectively address the early commitment issues arising from mainstream top-down approaches. We showed that the proposed coarse-pose proposal filtering can be applied to mainstream top-down human detectors and can consistently improve pose estimation performance. Experiments on COCO, and OCHuman datasets show the efficacy of the proposed approach.

Future work includes the improvement of human detection and pose estimation performances when trained with less data or noisy annotations. One issue that lacks considerations in mainstream object detectors is that a large number of detection proposals must be kept for consideration in the processing pipeline, while a large portion of these proposals is not valid. Our coarse-pose proposal filtering achieves some level of proposal refinement and thus alleviates this problem; however potential improvements can be further explored. Finally, the proposed hybrid framework can be extended to improve other related problems such as the *human parsing*, where the *instance segmentation* of each individual and each body part can be improved using a similar hybrid framework.

Acknowledgment

This work is partly supported by US National Science Foundation Research Grant #IIS-2008532.

Fig. 10: **Human detection and pose estimation results (part I)**: (1^{st} row) large scale variations of the individuals appearing in the scene and large viewpoint variations, (2^{nd} row) heavy keypoint occlusions, (3^{rd} row) highly overlapping body layouts among interactions or group activities.

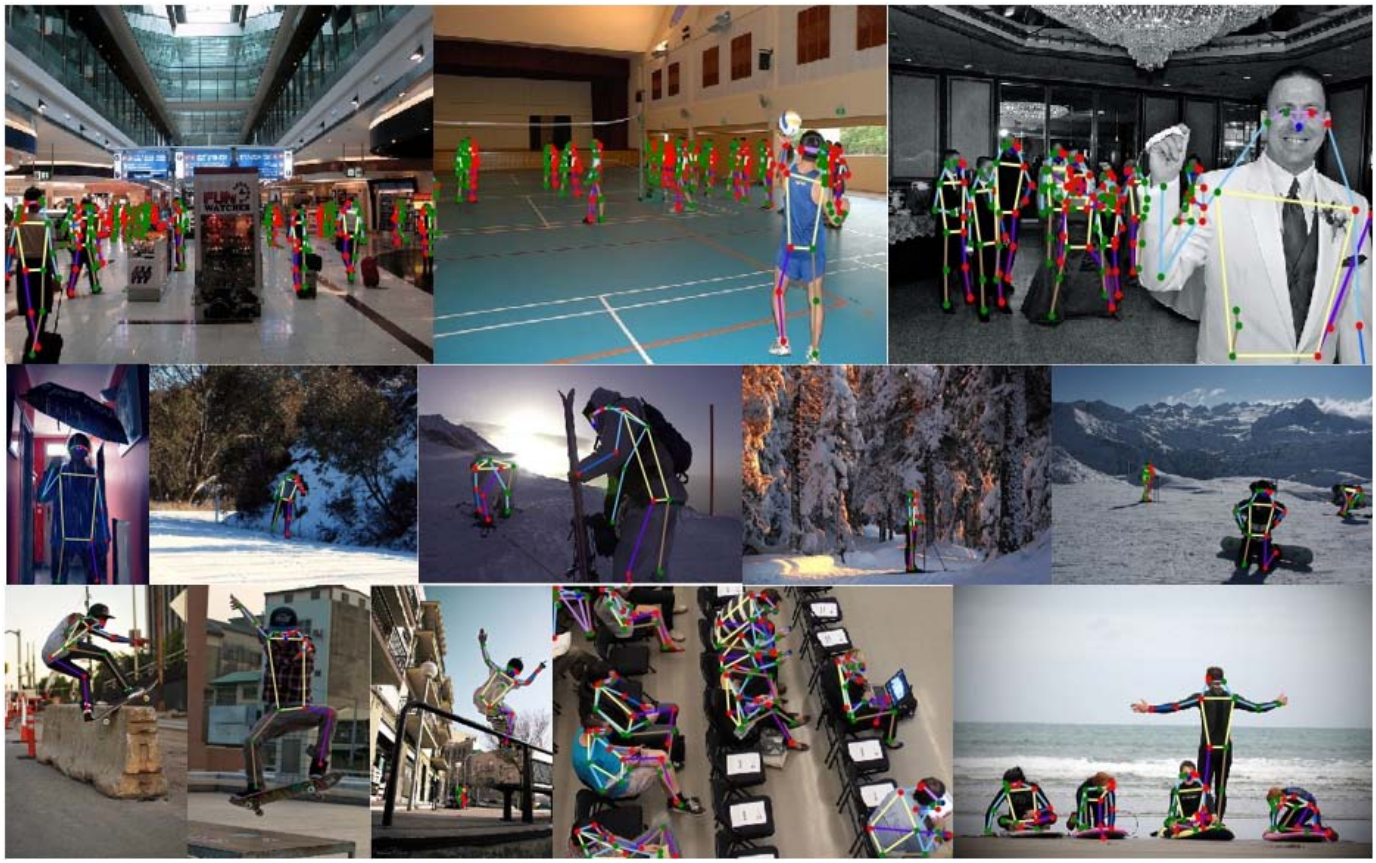

Fig. 11: **Human detection and pose estimation results (part II)**: (1^{st} row) crowded scenes in an open venue, (2^{nd} row) rich illumination and appearance vairations, (3^{rd} row) diverse activities with distinct body layouts and movements.

Fig. 12: **Failure or imperfect cases** of DetPoseNet. (1^{st} row) partial human body, where a scarce set of body keypoints are confidently detected, however the major portion of the human body is not visible, causing erroneous whole body regression results. (2^{nd} row) false body keypoint detection or regression results. (3^{rd} row) miss-detection of body keypoints that causes the miss-detection of the whole person or wrong body regression results. We note that DetPoseNet is not designed to handle these cases. Many of these difficulties can be be handled with extra consideration or model sophistication; however these is not the main focus of this paper.

REFERENCES

[1] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE PAMI*, vol. 35, no. 12, pp. 2878–2890, 2012.

[2] G. Gkioxari, B. Hariharan, R. B. Girshick, and J. Malik, "Using k-Poselets for detecting people and localizing their keypoints," in *CVPR*, 2014, pp. 3582–3589.

[3] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *ICCV*, 2011, pp. 723–730.

[4] U. Iqbal and J. Gall, "Multi-person pose estimation with local joint-to-person associations," in *ECCV*, 2016, pp. 627–642.

[5] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. P. Murphy, "Towards accurate multi-person pose estimation in the wild," in *CVPR*, 2017, pp. 3711–3719.

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017, pp. 2334–2343.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017, pp. 2961–2969.

[8] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *CVPR*, 2018, pp. 7103–7112.

[9] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *ECCV*, 2018, pp. 472–487.

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299.

[12] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *ECCV*, 2018, pp. 269–286.

[13] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *NeurIPS*, 2017, pp. 2277–2287.

[14] M. Kocabas, S. Karagoz, and E. Akbas, "MultiPoseNet: Fast multi-person pose estimation using pose residual network," *ECCV*, pp. 437–453, 2018.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

[16] S.-H. Zhang, R. Li, X. Dong, P. L. Rosin, Z. Cai, H. Xi, D. Yang, H.-Z. Huang, and S. Hu, "Pose2Seg: Detection free human instance segmentation," in *CVPR*, 2019.

[17] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *ECCV*, 2018, pp. 731–746.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE PAMI*, vol. 37, no. 9, pp. 1904–1916, 2015.

[20] R. Girshick, "Fast R-CNN," in *ICCV*, 2015, pp. 1440–1448.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.

[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.
[23] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *IJCV*, vol. 61, no. 1, pp. 55–79, 2005.
[24] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Strike a pose: tracking people by finding stylized poses," in *CVPR*, 2005, pp. 271–278.
[25] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *CVPR*, 2010, pp. 623–630.
[26] A. Mykhaylo, R. Stefan, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009, pp. 1014–1021.
[27] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *CVPR*, 2013, pp. 588–595.
[28] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation." in *BMVC*, 2010.
[29] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *ECCV*, 2008, pp. 710–724.
[30] L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *CVPR*, 2006, pp. 2041–2048.
[31] X. Lan and D. P. Huttenlocher, "Beyond trees: Common-factor models for 2D human pose recovery," in *ICCV*, 2005, pp. 470–477.
[32] L. Karlinsky and S. Ullman, "Using linking features in learning non-parametric part models," in *ECCV*, 2012, pp. 326–339.
[33] M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *CVPR*, 2013, pp. 3041–3048.
[34] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016, pp. 483–499.
[35] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016, pp. 4724–4732.
[36] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *CVPR*, 2014, pp. 2329–2336.
[37] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *CVPR*, 2015, pp. 648–656.
[38] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NeurIPS*, 2014, pp. 1799–1807.
[39] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *NeurIPS*, 2014, pp. 1736–1744.
[40] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *CVPR*, 2014, pp. 1653–1660.
[41] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *FG*, 2017, pp. 468–475.
[42] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *ECCV*. Springer, 2016, pp. 717–732.
[43] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *CVPR*, 2017, pp. 1831–1840.
[44] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *ICCV*, 2017, pp. 1281–1290.
[45] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial PoseNet: A structure-aware convolutional network for human pose estimation," in *ICCV*, 2017, pp. 1212–1221.
[46] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *ECCV*, 2018, pp. 190–206.
[47] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *arXiv:1901.00148*, 2019.
[48] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," in *ICCV*, 2015, pp. 1913–1921.
[49] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *ECCV*, 2014, pp. 33–47.
[50] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," *Wiley-IEEE Press*, pp. 237–243, 2001.
[51] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010, pp. 249–256.
[52] Y. Bengio, P. Y. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.
[53] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," *IEEE PAMI*, vol. 43, pp. 172–186, 2021.
[54] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *CVPR*, 2016, pp. 4929–4937.
[55] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *ECCV*, 2016, pp. 34–50.
[56] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "ArtTrack: Articulated multi-person tracking in the wild," in *CVPR*, 2017, pp. 1293–1301.
[57] X. Nie, J. Feng, J. Xing, and S. Yan, "Pose partition networks for multi-person pose estimation," in *ECCV*, 2018, pp. 705–720.
[58] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.
[59] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014, pp. 3686–3693.
[60] W. McNally, K. Vats, A. Wong, and J. McPhee, "Evopose2d: Pushing the boundaries of 2d human pose estimation using neuroevolution," *arXiv preprint arXiv:2011.08446*, 2020.

Lipeng Ke received the B.Eng. degree in Information Engineering from the China University of Mining and Technology, Xuzhou, China in 2015 and the M.S. degree in computer science from University of Chinese Academy of Sciences, Beijing, China in 2018. He is currently pursuing the Ph.D. degree with Department of Computer Science and Engineering, University at Buffalo, State University of New York, Buffalo, NY, USA.

His current research interests include pose estimation, object detection/tracking and deep learning.

Ming-Ching Chang is an Assistant Professor at the Department of Computer Science, College of Engineering and Applied Sciences (CEAS), University at Albany, State University of New York (SUNY). He was with the Department of Electrical and Computer Engineering from 2016 to 2018. He was an Adjunct Professor with the Computer Science Department from 2012-2016. During 2008-2016, he was a Computer Scientist at GE Global Research Center. He received his Ph.D. degree in the Laboratory for Engineering Man/Machine Systems (LEMS), School of Engineering, Brown University in 2008. He was an Assistant Researcher at the Mechanical Industry Research Labs, Industrial Technology Research Institute (ITRI) at Taiwan from 1996 to 1998. He received his M.S. degree in Computer Science and Information Engineering (CSIE) in 1998 and B.S. degree in Civil Engineering in 1996, both from National Taiwan University.

Dr. Chang's expertise includes video analytics, computer vision, image processing, and artificial intelligence. His research projects are funded by GE Global Research, IARPA, DARPA, NIJ, VA, and UAlbany. He is the recipient of the IEEE Advanced Video and Signal-based Surveillance (AVSS) 2011 Best Paper Award - Runner-Up, the IEEE Workshop on the Applications of Computer Vision (WACV) 2012 Best Student Paper Award, the GE Belief - Stay Lean and Go Fast Management Award in 2015, and the IEEE Smart World NVIDIA AI City Challenge 2017 Honorary Mention Award. Dr. Chang serves as Co-Chair of the annual AI City Challenge CVPR 2018-2021 Workshop, Co-Chair of the IEEE Lower Power Computer Vision (LPCV) Annual Contest and Workshop 2019-2021, Program Chair of the IEEE Advanced Video and Signal-based Surveillance (AVSS) 2019, Co-Chair of the IWT4S 2017-2019, Area Chair of IEEE ICIP (2017, 2019-2021) and ICME (2021). He has authored more than 100 peer-reviewed journal and conference publications, 7 US patents and 15 disclosures. He is a senior member of IEEE.

Honggang Qi received the M.S. degree in computer science from Northeast University, Shenyang, China, in 2002, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008.

He is currently an Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. His current research interests include computer vision, video coding and very large scale integration design.

Siwei Lyu is currently an Professor with the Computer Science Department, University at Buffalo, State University of New York. He received the B.S. degree in information science and the M.S. degree in computer science from Beijing University, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree in computer science from Dartmouth College, Hanover, NH, USA, in 2005. He has authored one book, one book chapter, and over 140 refereed journal and conference papers. His current research interests include digital image forensics, computer vision, and machine learning. Dr. Lyu was a recipient of the IEEE Signal Processing Society Best Paper Award in 2011 and the U.S. NSF CAREER Award in 2010.