The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

PoseHD: Boosting Human Detectors Using Human Pose Information

Zhijian Liu,[†] Bowen Pan,[†] Yuliang Xiu,[†] Cewu Lu[‡*]
Shanghai Jiao Tong University, Shanghai, China 200240
[†]{liuzhijian, googletornado, yuliangxiu}@sjtu.edu.cn, [‡]lu-cw@cs.sjtu.edu.cn

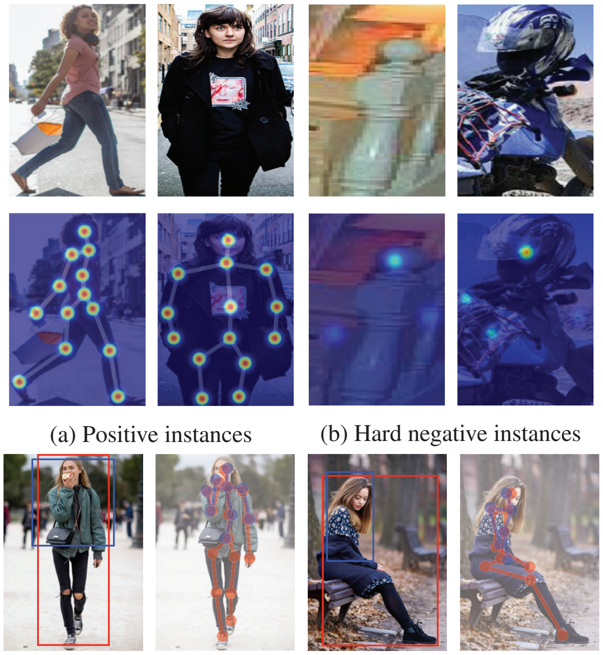
Abstract

As most recently proposed methods for human detection have achieved a sufficiently high *recall rate* within a reasonable number of proposals, in this paper, we mainly focus on how to improve the *precision rate* of human detectors. In order to address the two main challenges in precision improvement, *i.e.*, i) *hard background instances* and ii) *redundant partial proposals*, we propose the novel PoseHD framework, a top-down pose-based approach on the basis of an arbitrary state-of-the-art human detector. In our proposed PoseHD framework, we first make use of human pose estimation (in a batch manner) and present pose heatmap classification (by a convolutional neural network) to eliminate hard negatives by extracting the more detailed structural information; then, we utilize pose-based proposal clustering and reranking modules, filtering redundant partial proposals by comprehensively considering both holistic and part information. The experimental results on multiple pedestrian benchmark datasets validate that our proposed PoseHD framework can generally improve the overall performance of recent state-of-the-art human detectors (by 2-4% in both mAP and MR metrics). Moreover, our PoseHD framework can be easily extended to object detection with large-scale object part annotations. Finally, in this paper, we present extensive ablative analysis to compare our approach with these traditional bottom-up pose-based models and highlight the importance of our framework design decisions.

Introduction

Human detection, a long-standing task in computer vision, has been extensively studied due to its wide applications such as robotics, surveillance, and semantic understanding of video footage. There have been a number of recently proposed human detectors (Girshick et al. 2014; Girshick 2015; Ren et al. 2015; Redmon and Farhadi 2017; Liu et al. 2016; Zhang et al. 2016a), and most human detectors perform well in terms of their recall rates. However, these approaches normally suffer from low precision rates.

The challenges of improving precision are two-fold. The first is the *hard background instances* (hard negatives). Some non-human instances (such as the pillar and the helmet in Figure 1b) are identified as human because their overall appearance is human-like (containing some head-like and torso-like parts). It is impractical to collect all these examples as training data. The second challenge is *redundant partial proposals* (as illustrated in Figure 1c). These partial proposals are produced mainly because of the occlusion examples (where only part of the human body is visible) in the training data. In order to correctly detect these examples, the models are encouraged to also generate some partial proposals in their predictions. These proposals are too difficult to be eliminated by the non-maximum suppression (NMS).



(a) Positive instances

(b) Hard negative instances

(c) Redundant partial proposals (in blue box)

Figure 1: Two typical challenges of human detection: i) hard background instances (in Figure 1b), and ii) redundant partial proposals (in Figure 1c). In Figure 1a and Figure 1b, images in the first row are some example proposals, and those in the second row are the heatmaps containing all human keypoints. In Figure 1c, we include the human pose for each proposal.

In this paper, our core insight is to introduce pose information to human detectors in a top-down way. We review the mechanism of how we recognize a human: apart from overall

[*]Cewu Lu is the corresponding author.

appearance, we also employ more structural details of human posture to confirm our judgment. The benefits of incorporating pose prior to human detectors are two-fold. First, we can eliminate hard negatives by utilizing the structural information provided by human pose to verify whether a proposal is human. Second, the structural information gives measure of how much the human instance is covered, thereby penalizing the partial proposals.

Specifically, we extract the pose heatmaps for each proposal by a batch estimation algorithm and perform classification on these heatmaps to eliminate the hard negatives. Then, we group the proposals of the same human instances together by pose-based proposal clustering. We can then filter out partial proposals by removing anomalies from each cluster. Experimental results on multiple benchmarks validate that our PoseHD framework is able to improve the performance of human detectors.

Prior work has explored pose information in a bottom-up manner, whereas we take a top-down approach. Specifically, the previous *bottom-up frameworks* detect the human keypoints first, and then produce the proposals by grouping these keypoints together; however, our *top-down framework* predicts the proposals directly, and then employs the human keypoints to refine the prediction. The motivation for these two frameworks is entirely different: the bottom-up framework aims to improve the *recall rates* since the keypoints are much easier to detect than the proposals, while our top-down framework aims to improve the *precision rates* because the human pose provides extra information to remove the false positives. As recent human detectors achieve high recall rates, we believe that the bottom-up method (precision rates) should be studied more.

In summary, this paper has three major contributions:

1. We provide a new methodology for improving the human detectors by introducing pose information in a top-down manner.
2. We employ a novel combination of modules (pose estimation and classification) to filter out the hard negatives effectively.
3. We propose a novel technique (pose-based proposal clustering and reranking) to identify and eliminate the partial proposals.

Related Work

Our framework draws on recent work in human detection and part-based models (both with and without part annotations).

Human Detection Several pioneering approaches (Dollár et al. 2009; Dollar et al. 2014; Nam, Dollár, and Han 2014; Benenson et al. 2014) have demonstrated the effectiveness of the hand-crafted features, including the integrated channel feature (ICF). With the introduction of convolutional neural networks (CNN) to object detection, many recent human detectors (Benenson et al. 2014; Tian et al. 2015b; Tian, Yonglong et al. 2015) are based on state-of-the-art networks, such as R-CNN (Girshick et al. 2014). To our knowledge, most state-of-the-art human detectors (Hosang et al. 2015; Tian et al. 2015b; Tian, Yonglong et al. 2015; Cai, Saberian, and Vasconcelos 2015) are hybrid models that employ only hand-crafted features (Dollár et al. 2009; Dollar et al. 2014) and deeply learned features (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015). Our PoseHD framework, on the other hand, uses structural pose information to enhance the state-of-the-art.

Part-Based Human Detectors without Part Annotations One stream of part-based human detectors is the models without part annotations. These models are trained only using the annotations of human proposals, and among these models, the deformable part model (DPM) (Felzenszwalb, McAllester, and Ramanan 2008) is the most classic one: it captures a fixed number of part templates by the HOG feature (Dalal and Triggs 2005). Based on the DPM, a set of models with different geometry structures (Song et al. 2013; Lin et al. 2015) are investigated by researchers, and some other DPM-based models (Fidler et al. 2013; Song et al. 2011; Zhang et al. 2011; Mottaghi 2012; Khan et al. 2012) incorporate different features from HOG. With the help of deep learning, several higher level features are introduced to the human detectors to further improve the performance (Ouyang and Wang 2012; 2013; Tian et al. 2015a; Luo et al. 2014).

In summary, these part-based models (without part annotations) select latent human parts according to some energy function defined by the features (such as HOG) and the relative position and scale of different part proposals. However, without explicit part annotations, these models cannot make use of semantic information about human parts. Also, since there is no extra information provided to these models, it is difficult to extract the implicit structure of human keypoints, which is very important for eliminating the hard negatives and partial proposals.

Part-Based Human Detectors with Part Annotations Our PoseHD framework falls into this category, where the models are trained with the annotations of human parts. (Chen et al. 2014) demonstrates a typical pipeline for these models: separately train the detectors for each part and combine the parts together to obtain the human proposals. Several researchers also build their human detectors in this way (Mohan, Papageorgiou, and Poggio 2001; Mikolajczyk, Schmid, and Zisserman 2004; Wu and Nevatia 2005; Enzweiler et al. 2010; Bourdev and Malik 2009; Bourdev et al. 2010; Zhang et al. 2014). In particular, (Bourdev and Malik 2009) and (Bourdev et al. 2010) introduce the concept of poselets (a novel definition of human parts), detect poselets by sliding windows, employ the spatial context to rescore the poselets, and cluster the poselets together to form the human proposals. This methodology can be also applied to human segmentation. In (Popa and Sminchisescu 2015), they first predict some segmentation candidates, match them with a set of mask priors, and finally fuse these priors together to obtain the prediction.

In spite of sharing some similarities with the previous part-based models, our top-down approach can generally achieve a higher recall rate compared to these bottom-up frameworks. Furthermore, our objective of introducing the human parts is to improve the precision rates (reject the hard negatives and partial proposals) by extracting the more detailed struc-

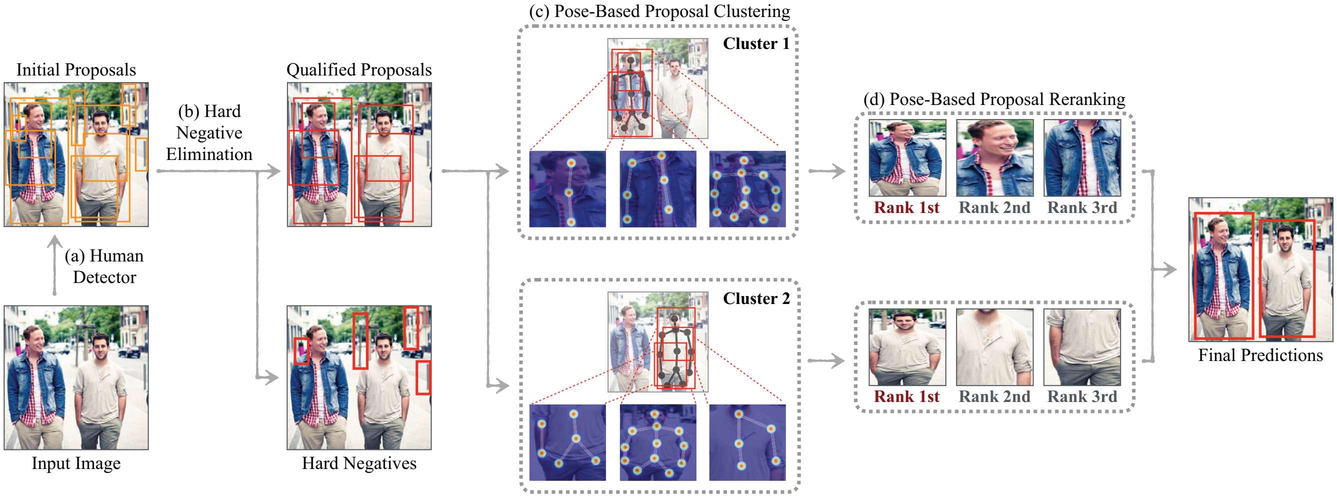


Figure 2: The illustration of our PoseHD framework. Given the initial proposals from human detector, our PoseHD framework makes use of the *pose estimation and classification* and *pose-based proposal clustering and reranking* modules to eliminate the *hard background instances* and *redundant partial proposal*, respectively.

tural information rather than improving the recall rates by addressing the problem of object occlusions.

Pose-Based Human Detection System

Our PoseHD framework is illustrated in Figure 2. Given an input image, we make use of an arbitrary human detector to generate initial proposals. Then, we apply pose estimation on these proposals (in a batch manner for efficiency) to produce a heatmap for each human keypoint. We perform binary classification on these pose heatmaps to eliminate false positives from initial proposals. Finally, we cluster each proposal based on keypoint location and rerank the proposals based on pose features to filter out partial proposals. Note that although we focus on human detection, each of our modules can be applied generically in other object categories.

Initial Human Proposals

We use an off-the-shelf human detector (Ren et al. 2015; Liu et al. 2016; Redmon and Farhadi 2017) for producing initial proposals. We demonstrate more details of how to modify these human detectors for our purposes in the EXPERIMENTS section.

Most state-of-the-art human detectors can achieve a sufficiently high recall rate within a reasonable number of proposals. For instance, on the PASCAL VOC 2007 (Everingham et al. 2010) dataset, FRCNN (Ren et al. 2015) achieves a recall rate of 97% within 100 proposals while SSD (Liu et al. 2016) achieves 95% with only 40 proposals. This inspires us to put our focus on improving the precision rate (eliminating false positives from the initial proposals) in order to further boost the overall performance of human detectors.

Hard Negative Elimination

We present a novel scheme to eliminate the hard background instances using pose information.

Pose Estimator Given the image of a proposal as input, a human pose estimator (Cao et al. 2017) outputs a set of pose heatmaps $\{H_p\}$. In the p^{th} pose heatmap, $H_p(x, y)$ represents the likelihood that (x, y) belongs to the p^{th} keypoint. We refer readers to (Cao et al. 2017) for more details. Note that we can easily substitute other pose estimators in our PoseHD framework.

True and False Positives Classification With the pose heatmaps, it is then possible to verify whether a proposal is a false positive by the structural information of human pose. The pose heatmaps of most false positives are similar to the examples in Figure 1b, satisfying one of the two patterns: i) for most keypoints, the pose heatmap is "cold" (low confidence) or ii) the relative position of keypoints violates the structural constraints of a human. Due to these patterns, we pose the elimination of false positives as a two-category (true positives and false positives) classification problem. Technically, we stack the pose heatmaps together as a multi-channel image and use the convolutional neural network to distinguish the false positives from true positives. More details can be found in the EXPERIMENTS section. After training the network, we eliminate the proposals that are classified as false positives.

Pose-Based Proposal Clustering and Reranking

After eliminating the hard negatives via classification, our goal is to filter out the partial proposals. For each ground-truth human instance, there is likely at least one sufficiently accurate (IoU ≥ 0.5) initial proposal, since our human detectors achieve high recall rates. In order to select one best proposal for each human instance, we follow a two-step framework. First, we determine which proposals correspond to the same human instance (*clustering*). Then, for each set of proposals, we select one optimal proposal with respect to some criterion (*ranking*).

In order to cluster the proposals, we must define a distance metric between pairs of proposals. With the human keypoints, our motivation is to measure the distance by the similarity between human poses of two proposals. We consider both a box-based metric and a pose-based metric; however we find both intuitively and empirically, the pose-based metric outperforms the box-based metric. In Figure 3, we can see that the pose-based distance can better distinguish between two distinct human instances with very close bounding boxes (by more semantic information of human instances).

Formally, let h_i and w_i denote the height and width of the i^{th} proposal, H_{ip} denote the heatmap for the p^{th} human keypoint, (x_{ip}, y_{ip}) denote the location with maximum value in H_{ip}, and v_{ip} denote the maximum value. The *pose-based distance* between the a^{th} and b^{th} proposal is defined as

$$\text{dist}_{\text{P}}(a,b) = \frac{\sum_p \min(v_{ap}, v_{bp}) \cdot \text{dist}_2(a,b,p)}{\sum_p \min(v_{ap}, v_{bp})}, \quad (1)$$

where $\text{dist}_2(a,b,p)$ is the normalized euclidean distance between the p^{th} keypoint of the a^{th} and b^{th} proposal:

$$\text{dist}_2(a,b,p) = \sqrt{\left(\frac{x_{ap} - x_{bp}}{h_a + h_b}\right)^2 + \left(\frac{y_{ap} - y_{bp}}{w_a + w_b}\right)^2}. \quad (2)$$

In Equation 1, the coefficients $\min(v_{ap}, v_{bp})$ are used to ensure that $\text{dist}_2(a,b,p)$ is taken into account if and only if the p^{th} keypoint appears in both proposals, and the denominator $\sum_p \min(v_{ap}, v_{bp})$ normalizes the distance by the number of keypoints that appear in both proposals. In Equation 2, the denominators $h_a + h_b$ and $w_a + w_b$ are used to normalize for the size of the proposal in each dimension.

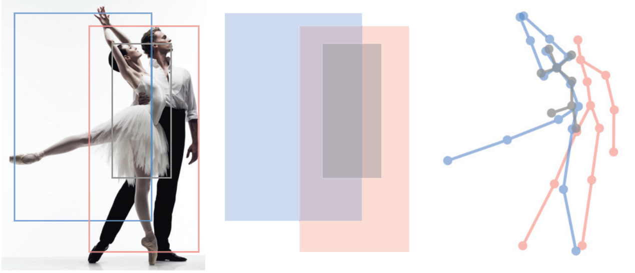
After we compute the pose-based distance between each pair of proposals, we perform the agglomerative hierarchical clustering (Larsen and Aone 1999) to cluster the proposals. We iteratively merge clusters until each pair of clusters exceeds a distance threshold D (a domain-specific hyperparameter). By properly setting the distance threshold D, we ensure that most clusters correspond to a unique human instance in the image.

Let C_k denote the k^{th} cluster. In order to determine the *representative proposal* r_k within the k^{th} cluster C_k, we select the one with the largest overlapping rate with the ground-truth human instances (denote as $o(\cdot)$) among all proposals p in this cluster C_k:

$$r_k = \arg\max_{p \in C_k} o(p) \iff \forall p \in C_k : o(r_k) \geq o(p). \quad (3)$$

Following this intuition, we pose the selection of representative proposals as a ranking problem (Joachims 2002). The setup is to design a feature vector $f(p)$ for each proposal p and to learn a weight vector θ such that the representative proposal r_k is expected to rank higher than all the other proposals p with respect to the ranking criterion $R(p) = \theta^{\text{T}} f(p)$:

$$\forall p \in C_k : R(r_k) = \theta^{\text{T}} f(r_k) \geq \theta^{\text{T}} f(p) = R(p). \quad (4)$$



(a) Proposals (b) Box-based view (c) Pose-based view

Figure 3: Clustering of proposals from box-based and pose-based perspective. From the box-based view (in Figure 3b), the proposal in black will clearly be assigned to the proposal in red since the black box is fully contained in the red box, while from the pose-based view (in Figure 3c), it will then be assigned to the proposal in blue, because its posture is highly overlapped with the blue one.

Another approach is to use linear regression to learn a weight vector θ such that $\theta^{\text{T}} f(p) \approx o(p)$. However, empirically, we find that learning the exact value of $o(p)$ is extremely hard, which, we conjecture, is because of the high variance of $f(p)$ and the highly non-linear relationship between $o(p)$ and $f(p)$. Since we are more concerned with maintaining the relative order of proposals, rather than precisely estimating the overlapping rate of each individual proposal, the ranking solution is more suited to our purposes than linear regression.

In this paper, we employ a four-dimensional feature for the i^{th} proposal, capturing both holistic and part information. The feature vector is composed of the following quantities:

- Confidence score given by the human detector.
- Normalized area size of the proposal: $h_i w_i / h_k^{\text{M}} w_k^{\text{M}}$. Here, h_i and w_i are the height and width of the proposal, while h_k^{M} and w_k^{M} (where $i \in C_k$) are the median height and median width of the proposals in the same cluster.
- Overall pose energy of the proposal: $\sum_p v_{ip}$. This value is essentially a soft count of the keypoints that have appeared in the proposal.
- Overlapping rate with the *average pose* of the cluster. Since the proposals in the same cluster always correspond to the same human instance, the *average pose* will then have fewer uncertainties, and it can roughly infer the region of human proposal. The *average pose* is computed by the pixel-wise average of heatmaps on the k^{th} cluster (where $i \in C_k$), which is

$$H_p^k(x,y) = \frac{1}{|C_k|} \sum_{j \in C_k} H_{jp}(x - x_j, y - y_j), \quad (5)$$

where (x_j, y_j) denotes the top-left corner of the j^{th} proposal. We extract the keypoints on H^k to form the *average pose*. Finally, the overlapping rate is computed by the intersection over union (IoU) between the i^{th} proposal and the bounding box of *average pose* of the k^{th} cluster.

For training the proposal reranking algorithm, we make use of the *LambdaMART* (Wu et al. 2010) implementation in RankLib, which solves the ranking problem by forests of boosted decision trees. After learning the weight vector θ, we select the representative proposal for each cluster by $R(p) = \theta^{\mathrm{T}} f(p)$, and remove all non-representative proposals in order to eliminate the partial proposals.

Discussion The non-maximum suppression (NMS) can be seen as a non-semantic version of our two-step approach: it utilizes an IoU-based distance for clustering in the first step, and in the second step, it employs the area size as the ranking criterion. In this sense, our framework is more general than the NMS algorithm and allows us to incorporate semantic information (human pose). Furthermore, several bottom-up models (Bourdev and Malik 2009; Bourdev et al. 2010) also employ clustering in their frameworks, but the motivation is entirely different: these bottom-up models use the clustering to group keypoints together to form the proposals, while our top-down framework employs the clustering to group proposals together to remove partial proposals.

Batch Estimation Algorithm for Acceleration

Even though (Cao et al. 2017) runs in real-time, the time overhead of calling the pose estimator for all initial proposals is still prohibitive. However, we leverage the fact that the human pose estimator is agnostic to the number of human instances and can therefore generate heatmaps for multiple proposals simultaneously. We propose the following batch pose estimation algorithm to reduce the number of calls to pose estimator. First, we group the proposals with IoU $\geq \gamma$ together by greedy merging. Then, we call the pose estimator on the tightest bounding box that encompasses all proposals in each group. Finally, the pose of each proposal is cropped out according to the area covered by the original proposal.

Our method reduces the number of calls to human pose estimator from once *per proposal* to once *per group*. We can also trade off between time and accuracy by increasing/decreasing γ. If γ is set to be 0, we call the pose estimator once for the entire image and crop the pose heatmaps for all proposals simultaneously. Although this is fast, the resulting poses are inaccurate since the recall rate of (Cao et al. 2017) is relatively low (see the ABLATION STUDY section). On the other hand, if γ is set to be $+\infty$, we revert to our original algorithm, calling to the pose estimator once per proposal.

Experiments

Datasets

The performance of PoseHD is evaluated across three pedestrian benchmark datasets: INRIA (Dalal and Triggs 2005), ETH (Ess, Leibe, and Van Gool 2007) and PASCAL VOC 2007 (Everingham et al. 2010). Note that for the PASCAL VOC 2007 dataset, we only use the data from human category during training and testing.

Evaluation Metrics

The IoU threshold used to determine the true positives is set to be 0.5, and the evaluation metrics we used are mean *Average Precision* (mAP) (Everingham et al. 2010) and log-average *Miss Rate on False Positive Per Image* in $[10^{-2}, 1]$ (MR) (Zhang et al. 2016b). These two metrics characterize the detection performance from two slightly different aspects, and their values might not be totally correlated.

Experiment Settings

In this section, we describe some experimental settings and details about generating data and training models.

Human Proposals with High Recall Rate We perform our experiments based on FRCNN (Ren et al. 2015), SSD (Liu et al. 2016) and YOLOv2 (Redmon and Farhadi 2017). In these human detectors, confidence score thresholding and non-maximum suppression (NMS) are normally used to reduce the number of output proposals; however, reducing the number of output proposals diminishes the recall rate. Therefore, we remove these two steps from our human detector, and instead, we generate a more reasonable number of proposals by selecting the top-k proposals (with respect to the confidence scores) without thresholding. Empirically, we found that $k = 100$ is a good setting for these baselines to produce proposals with high recall rate.

Hard Negative Elimination We make use of the trained real-time multi-person pose estimator (Cao et al. 2017) to produce the pose heatmaps, and in the batch estimation algorithm, we use a different IoU threshold γ on each dataset to ensure a constant number of calls to the pose estimator.

For the classification network, we use a slightly modified version of AlexNet (Krizhevsky, Sutskever, and Hinton 2012). We change the number of input channels of the first convolutional layer from 3 (RGB channels) to 18 (the number of human keypoints) and resize the pose heatmaps to 224×224 in order to fit the input size of AlexNet.

Because the proposals from human detectors are usually limited in size and biased to one category (most proposals are false positives), we augment the dataset by generating some examples based on the annotations: a large number of proposals are sampled around the ground-truth human instances, and then each proposal is labeled as true or false positive if its overlapping rate (with ground-truth human instances) is large or small respectively. During training, we horizontally flip each proposal to double the size of our dataset.

We use TensorFlow (Abadi et al. 2016) and Nvidia Titan X GPU to train the classification network. We initialize the weights randomly from scratch, and the optimization is carried out using Adam (Kingma and Ba 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use a fixed learning rate of $\gamma = 10^{-3}$ and mini-batch size of 128.

Pose-Based Proposal Clustering and Reranking For the proposal clustering, we set a different distance threshold D for each human detector and each dataset. In our experiments, we only use part of the features in *LambdaMART* since we are only scoring the 1/0 loss on the representative proposals, rather than scoring the ranked results by an IR-style metric. Similar to the training of classification network, we also augment the dataset with some generated examples: we produce

	INRIA		ETH		PASCAL VOC 2007	
Model	mAP	MR	mAP	MR	mAP	MR
FRCNN (Ren et al. 2015)	88.8	12.1	66.0	62.7	75.8	30.8
Pose-FRCNN	90.3	8.9	69.8	58.1	78.8	27.8
Improvement	**1.5**	**3.2**	**3.8**	**4.6**	**3.0**	**3.0**
SSD (Liu et al. 2016)	82.0	14.6	67.8	56.1	76.7	30.0
Pose-SSD	86.1	10.9	69.6	52.5	80.7	25.9
Improvement	**4.1**	**3.7**	**1.8**	**3.6**	**4.0**	**4.1**
YOLOv2 (Redmon and Farhadi 2017)	88.9	12.0	59.5	70.6	79.1	27.4
Pose-YOLOv2	90.4	8.9	61.5	67.4	80.3	25.1
Improvement	**1.5**	**3.1**	**2.0**	**3.2**	**1.2**	**2.3**
Average improvement	**2.4**	**3.3**	**2.6**	**3.8**	**2.7**	**3.1**

Table 1: Human detection performance on three benchmark datasets (Dalal and Triggs 2005; Ess, Leibe, and Van Gool 2007; Everingham et al. 2010). In this table, rows in gray contain results by applying the PoseHD on top of the corresponding baseline.

a cluster of proposals by sampling several bounding boxes near one ground-truth human instance.

Results

We present the performance on multiple pedestrian benchmark datasets in Table 1, where all baseline detectors follow their default settings in their original papers. Our proposed PoseHD framework improves on these strong baselines across several datasets, and in average, we observe a 2-4% performance gain in both mAP and MR metrics.

As observed in Table 1, our PoseHD framework is fairly effective for human detectors which can produce proposals with high recall rate (such as FRCNN and SSD), while the improvements achieved on these with lower recall (such as YOLO) are relatively smaller. We conjecture that this might be because YOLO mainly suffers from the localization of bounding boxes and gives fewer predictions of background instances than the other two approaches (a similar conclusion has been drawn by (Redmon et al. 2016)).

Discussion

As some current detectors aim at real-time performance, we briefly analysis the tradeoff between the 2-4% performance gain and the computational overhead of PoseHD framework. Because the inference time of AlexNet is small enough to be omitted, the extra computation mainly falls in detecting human pose for each initial proposal, which by batch estimation algorithm, requires constant times of human pose estimation. Since (Cao et al. 2017) can be run at real-time, and all predictions can be run in parallel, the time overhead is relatively minor. In addition, because optimizing pose estimation is an active area of research, we expect that the overhead of running pose estimation in our framework will be small enough for real-time human detection in the near future.

Our PoseHD framework can be extended to object detection if provided with a large-scale object dataset with part annotations, since the concept of human keypoints is a subset of object parts. However, the number of part annotations for most object categories (*e.g.* in the PASCAL-Part dataset) is still very limited, and therefore, we intend to include this in future research.

Ablation Study

In order to highlight the importance of our design decisions in the PoseHD framework, we present the results of several ablation analysis. We provide a quantitative comparison between the top-down and bottom-up frameworks to distinguish our approach from these traditional pose-based models. All experiments in this section are conducted based on the INRIA (Dalal and Triggs 2005) dataset with SSD as the baseline human detector.

Effectiveness of Building Blocks

Our PoseHD framework has two building blocks: i) hard negative elimination (HNE), and ii) pose-based proposal clustering and reranking (PPCR). They are proposed to eliminate hard negatives and partial proposals, respectively.

We conduct ablation analysis by removing these two modules from the PoseHD framework, and present the results in Table 2. Note that, after removing the PPCR from PoseHD, we apply the NMS algorithm with IoU threshold 0.5 instead for fair comparison.

From Table 2, both HNE and PPCR are indispensable for our PoseHD framework, each of which gives a roughly 2% boost in mAP, and their effects are orthogonal to each other.

Moreover, the PPCR alone can be seen as an extension of the NMS algorithm. From this table, the NMS algorithm achieves a reduction from 100 to 6 proposals while the PPCR gives a reduction from 100 to 3 (with roughly the same recall

Model	Proposal Number	Recall Rate	mAP
Pose-SSD	2.9 / image	98.0%	**86.1**
without HNE	3.0 / image	99.3%	83.6
without PPCR	5.6 / image	98.1%	84.4
Original SSD	6.0 / image	99.8%	82.0

Table 2: Ablation analysis of building blocks of the PoseHD framework. In this table, HNE represents *Hard Negative Elimination* and PPCR represents *Pose-Based Proposal Clustering and Reranking*.

Model	mAP
Pose-SSD	86.1
without confidence score by human detector	**84.4**
without normalized area size of proposal	85.6
without overall pose energy of proposal	84.7
without overlapping rate with average pose	**83.5**
Original SSD	82.0
Oracle Pose-SSD	86.8

Table 3: Ablation analysis of different features for proposal reranking. In this table, Oracle Pose-SSD represents the modified Pose-SSD system (with the knowledge of overlapping rate during proposal reranking).

rate). In the supplementary material, we also showcase some examples that compare the predictions of NMS and PPCR. In these examples, the PPCR generally removes most partial proposals from initial prediction, while in some other cases (the last example in the second row), the partial proposals are not entirely removed mainly because our keypoint-based distance metric is not resilient to inaccurate pose predictions. However, with the further research on human pose estimator, we believe that this issue will be gradually addressed.

Effectiveness of Different Features

During the proposal reranking, we compute four features to train the ranking function. To highlight their effectiveness, we conduct ablative analysis by removing each feature from the framework. In order to show the potential of designing features, we also construct the Oracle PoseHD framework, which directly uses the overlapping rate as ranking function.

Table 3 suggests that *overlapping rate with average pose* and *confidence score by human detector* are the most important classes of features since they provide a sense of how a proposal looks like human from both holistic and part view. Also, the Oracle Pose-SSD achieves a 1% gain compared to the original Pose-SSD, which means that there is still much room for feature design.

Model	Recall Rate	mAP	MR
Pose-FRCNN	**98.1%**	90.3	**8.9**
Pose-YOLOv2	98.0%	**90.4**	**8.9**
Pose-SSD	98.0%	86.1	10.9
(Cao et al. 2017)	92.5%	83.5	32.0

Table 4: Comparison between the top-down and bottom-up frameworks. In this table, (Cao et al. 2017) represents the bottom-up human detectors adapted from Cao et al..

Comparison with Bottom-Up Models

Our PoseHD framework is designed in a *top-down* manner while most related work (Bourdev et al. 2010; Enzweiler et al. 2010; Chen et al. 2014) falls into the *bottom-up* framework (two types of frameworks are explained in Introduction). Since most bottom-up models do not even use deep learning, it is impractical to plug them in the state-of-the-art systems. On the other hand, it is unfair to compare with their original results (most of them only achieve roughly 50 in mAP on the PASCAL VOC dataset).

To quantitatively compare these two kinds of frameworks, we adapt the (Cao et al. 2017) into a bottom-up human detector. In detail, we compute bounding box of keypoints for each human pose, and train a linear regressor to finetune the bounding boxes (since the bounding box of keypoints is typically slightly different from the human proposal). This system can be seen as an upgraded version of previous bottom-up models since in (Cao et al. 2017), they first detect human keypoints, and then group them into different human poses.

From Table 4, the PoseHD systems significantly outperform the modified (Cao et al. 2017) in both mAP and MR metrics, which is because the modified (Cao et al. 2017) has a comparatively low recall rate. This result gives us a sense that top-down framework is more appealing when the recall rate is already high enough, and we believe these two frameworks will indeed complement each other in precision and recall improvement.

Conclusion

In this paper, we summarized two main challenges of current human detectors: hard background instances and redundant partial proposals, and we proposed a novel PoseHD framework, a top-down pose-based approach with several effective techniques. Our PoseHD framework is generic enough to be plugged into any existing human detector, and the experimental results on several benchmarks suggest that it is able to boost the overall performance of many human detectors. We also presented analytical and quantitive comparisons between bottom-up pose-based models and our top-down framework.

There are multiple aspects in our current framework that we intend to include in our future work: i) improving the localization of proposals with pose information and ii) extending our framework to object detection. We believe that our methodology will help advance the state-of-the-art in object detection.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P. A.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; and Zheng, X. 2016. TensorFlow - A System for Large-Scale Machine Learning. In *OSDI*.

Benenson, R.; Omran, M.; Hosang, J.; and Schiele, B. 2014. Ten Years of Pedestrian Detection, What Have We Learned? In *ECCV Workshop*.

Bourdev, L., and Malik, J. 2009. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*.

Bourdev, L. D.; Maji, S.; Brox, T.; and Malik, J. 2010. Detecting People Using Mutually Consistent Poselet Activations. In *ECCV*.

Cai, Z.; Saberian, M.; and Vasconcelos, N. 2015. Learning Complexity-Aware Cascades for Deep Pedestrian Detection. In *ICCV*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R.; and Yuille, A. L. 2014. Detect What You Can - Detecting and Representing Objects using Holistic Models and Body Parts. In *CVPR*.

Dalal, N., and Triggs, B. 2005. Histograms of Oriented Gradients for Human Detection. In *ICCV*.

Dollár, P.; Tu, Z.; Perona, P.; and Belongie, S. J. 2009. Integral Channel Features. *BMVC*.

Dollar, P.; Appel, R.; Belongie, S.; and Perona, P. 2014. Fast Feature Pyramids for Object Detection. *TPAMI*.

Enzweiler, M.; Eigenstetter, A.; Schiele, B.; and Gavrila, D. M. 2010. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*.

Ess, A.; Leibe, B.; and Van Gool, L. J. 2007. Depth and Appearance for Mobile Scene Analysis. In *ICCV*.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*.

Felzenszwalb, P.; McAllester, D.; and Ramanan, D. 2008. A Discriminatively Trained, Multiscale, Deformable Part Model. In *CVPR*.

Fidler, S.; Mottaghi, R.; Yuille, A.; and Urtasun, R. 2013. Bottom-Up Segmentation for Top-Down Detection. In *CVPR*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *CVPR*.

Girshick, R. 2015. Fast R-CNN. In *ICCV*.

Hosang, J.; Omran, M.; Benenson, R.; and Schiele, B. 2015. Taking a deeper look at pedestrians. In *CVPR*.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *KDD*.

Khan, F. S.; Anwer, R. M.; van de Weijer, J.; Bagdanov, A. D.; Vanrell, M.; and López, A. M. 2012. Color attributes for object detection. In *CVPR*.

Kingma, D., and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*.

Larsen, B., and Aone, C. 1999. Fast and effective text mining using linear-time document clustering. In *KDD*.

Lin, L.; Wang, X.; Yang, W.; and Lai, J.-H. 2015. Discriminatively Trained And-Or Graph Models for Object Shape Detection. *TPAMI*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.-Y.; and Berg, A. C. 2016. SSD - Single Shot MultiBox Detector. In *ECCV*.

Luo, P.; Tian, Y.; Wang, X.; and Tang, X. 2014. Switchable Deep Network for Pedestrian Detection. In *CVPR*.

Mikolajczyk, K.; Schmid, C.; and Zisserman, A. 2004. Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. In *ECCV*.

Mohan, A.; Papageorgiou, C.; and Poggio, T. 2001. Example-based object detection in images by components. *TPAMI*.

Mottaghi, R. 2012. Augmenting deformable part models with irregular-shaped object patches. In *CVPR*.

Nam, W.; Dollár, P.; and Han, J. H. 2014. Local Decorrelation For Improved Pedestrian Detection. In *NIPS*.

Ouyang, W., and Wang, X. 2012. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*.

Ouyang, W., and Wang, X. 2013. Joint Deep Learning for Pedestrian Detection. In *ICCV*.

Popa, A.-I., and Sminchisescu, C. 2015. Parametric Image Segmentation of Humans with Structural Shape Priors. *arxiv*.

Redmon, J., and Farhadi, A. 2017. YOLO9000: Better, Faster, Stronger. In *CVPR*.

Redmon, J.; Divvala, S. K.; Girshick, R. B.; and Farhadi, A. 2016. You Only Look Once - Unified, Real-Time Object Detection. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*.

Simonyan, K., and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Song, Z.; Chen, Q.; Huang, Z.; Hua, Y.; and Yan, S. 2011. Contextualizing object detection and classification. In *CVPR*.

Song, X.; Wu, T.; Jia, Y.; and Zhu, S.-C. 2013. Discriminatively Trained And-Or Tree Models for Object Detection. In *CVPR*.

Tian, Y.; Luo, P.; Wang, X.; and Tang, X. 2015a. Deep Learning Strong Parts for Pedestrian Detection. In *ICCV*.

Tian, Y.; Luo, P.; Wang, X.; and Tang, X. 2015b. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*.

Tian, Yonglong; Luo, Ping; Wang, Xiaogang; and Tang, Xiaoou. 2015. Deep Learning Strong Parts for Pedestrian Detection. In *ICCV*.

Wu, B., and Nevatia, R. 2005. Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors. In *ICCV*.

Wu, Q.; Burges, C. J. C.; Svore, K. M.; and Gao, J. 2010. Adapting boosting for information retrieval measures. *IR*.

Zhang, J.; Huang, K.; Yu, Y.; and Tan, T. 2011. Boosted local structured HOG-LBP for object localization. In *CVPR*.

Zhang, N.; Donahue, J.; Girshick, R.; and Darrell, T. 2014. Part-Based R-CNNs for Fine-Grained Category Detection. In *ECCV*.

Zhang, L.; Lin, L.; Liang, X.; and He, K. 2016a. Is Faster R-CNN Doing Well for Pedestrian Detection? In *ECCV*.

Zhang, S.; Benenson, R.; Omran, M.; Hosang, J. H.; and Schiele, B. 2016b. How Far are We from Solving Pedestrian Detection? In *CVPR*.