Generative Partition Networks for Multi-Person Pose Estimation

Xuecheng Nie\(^1\), Jiashi Feng\(^1\), Junliang Xing\(^2\), Shuicheng Yan\(^1,3\)
\(^1\)ECE Department, National University of Singapore, Singapore
\(^2\)Institute of Automation, Chinese Academy of Sciences, Beijing, China
\(^3\)Qihoo 360 AI Institute, Beijing, China
niexuecheng@u.nus.edu, elefjia@nus.edu.sg, jlxing@nlpr.ia.ac.cn, yanshuicheng@360.cn

Abstract

This paper proposes a new framework, named Generative Partition Network (GPN), for addressing the challenging multi-person pose estimation problem. Different from existing pure top-down and bottom-up solutions, the proposed GPN models the multi-person partition detection as a generative process from joint candidates and infers joint configurations for person instances from each person partition locally, resulting in both low joint detection and joint partition complexities. In particular, GPN designs a generative model based on the Generalized Hough Transform framework to detect person partitions via votes from joint candidates in the Hough space, parameterized by centroids of persons. Such generative model produces joint candidates and their corresponding person partitions by performing only one pass of joint detection. In addition, GPN formulates the inference procedure for joint configurations of human poses as a graph partition problem and optimizes it locally. Inspired by recent success of deep learning techniques for human pose estimation, GPN designs a multi-stage convolutional neural network with feature pyramid branch to jointly learn joint confidence maps and Hough transformation maps. Extensive experiments on two benchmarks demonstrate the efficiency and effectiveness of the proposed GPN.

1 Introduction

Multi-person pose estimation aims to localize body joints of multiple persons in a 2D monocular image. Despite long history of prior investigation, this problem remains very challenging due to complex joint configuration, partial or even complete joint occlusion, significant overlapping between neighboring persons, unknown number of persons and difficult joint allocation to multiple persons. These challenges feature the unique property of multi-person pose estimation compared with the single-person setting. Existing approaches usually perform joint detection and joint partition (via inference) separately with different strategies to solve the problems, deriving two solution techniques: the top-down approaches \(^1\) first generate person partitions and then apply single-person pose estimators to the partitions; and the bottom-up approaches \(^6\) generate joint candidates at first and then partition them into person instances by affinities from articulated kinematics of human pose.

The top-down approaches are the most straightforward strategy for multi-person pose estimation, which directly leverage existing techniques for person detection and single-person pose estimation, and have low joint partition complexity. However, their performance is critically limited by the quality of the generated partitions from person detectors. If the employed person detector fails to detect a person instance (due to some distracting factors, e.g., occlusion and overlapping), there is no chance to recover these failure cases. In addition, their joint detection complexity is usually very high—linearly increasing with the number of persons—because they need to run the single-person pose estimator for each of the person partition sequentially.
Figure 1: Overview of the proposed Generative Partition Network for multi-person pose estimation. GPN takes an image, as shown in (a), as input for a convolutional neural network to predict the Hough transformation maps and joint confidence maps, shown in top and bottom of (b), respectively. Joint candidates can vote for the centroid hypotheses of multiple persons, as shown in (c), which results in a generative process for detecting person partitions, as shown in (d), via finding peaks in the Hough space. GPN infers joint configurations for person instances in each person partition locally. The final pose estimation result for multiple persons is shown in (e).

In contrast, the bottom-up approaches employ joint detectors to detect all the candidate joints at first, resulting in robustness to error from early commitment. Following such bottom-up strategy, joint candidates can be detected from the whole image by only running joint detector for once. Hence, the joint detection complexity of bottom-up approaches is generally lower than top-down approaches. However, they suffer from very high joint partition complexity. The inference procedure aims to allocate all the joint candidates to the corresponding persons, which usually involves solving NP-hard graph partition problems on densely connected graphs covering the whole images.

In this paper, we propose the Generative Partition Network (GPN) to overcome the limitations of the above two types of approaches. As shown in Figure 3, GPN solves multi-person pose estimation by (1) modeling the person detection and partition as a generative process inferred from joint candidates and (2) performing local inference to solve joint configurations for person instances. In particular, GPN adopts a generative model based on the Generalized Hough Transform framework to detect person partitions via votes from joint candidates in the Hough space, parameterized by the centroid of person. This generative model can produce joint candidates and their corresponding person partitions by employing joint detector for only once, thus reducing the joint detection complexity and overcoming the major drawback of top-down solution technique. In addition, GPN formulates the inference procedure for joint configurations of human poses as a graph partition problem and optimizes it in each person partition locally. The local optimization strategy can reduce the search space for finding optimal joint configuration, avoiding the high joint partition complexity challenging the bottom-up solution technique. Inspired by the recent success of deep learning techniques for human pose estimation [12, 13, 8] and object detection [14], GPN designs a multi-stage convolutional neural network architecture with a feature pyramid branch to simultaneously learn the joint confidence maps and Hough transformation maps with enhanced efficiency. Extensive experiments on MPII Multi-Person Pose and WAF benchmarks evidently show the efficiency and effectiveness of GPN.

Related works In this section, we simply review recent representative approaches for multi-person pose estimation. Top-down approaches: Iqbal and Gall [1] follow the top-down strategy by combining Faster-RCNN [15] based person detector with convolutional pose machine [13] based joint estimator. Later, Fang et al. [2] adapt spatial transformer network to improve the quality of person partitions and use hourglass network [16] for detecting joints in each person partition. Bottom-up approaches: Pishchulin et al. [6] and Insafutdinov et al. [7] follow the bottom-up strategy by exploring geometric and appearance constraints among joint candidates to associate them, and formulate the multi-person pose estimation as subset partition and labeling problems optimized by integer linear programming. Cao et al. [8] further propose part affinity fields, encoding location and orientation of limbs, to measure the confidence for each pair of joint candidates from the same person, and formulate joint partition as a bipartite matching problem. The above approaches however cannot overcome inherent limitations of top-down and bottom-up strategies as mentioned above.
2 Methodology

2.1 Our model

In this section, we introduce the proposed Generative Partition Network (GPN) model to solve the multi-person pose estimation problem. Throughout the paper, we use following notations. Let I denote an image containing multiple persons, \( p = \{p_1, p_2, \ldots, p_N\} \) denote the spatial coordinates of \( N \) joint candidates from all persons in I with \( p_i = (x_i, y_i)^\top \), \( \forall i = 1, 2, \ldots, N \), and \( u = \{u_1, u_2, \ldots, u_N\} \) denote the labels of body joints, \( u_i \in \{1, 2, \ldots, K\} \) with \( K \) the number of joint categories. For allocating joints via local inference, we also consider the proximities between joints which are denoted as \( b \in \mathbb{R}^{N \times N} \). Here \( b_{u,v} \) indicates the proximity between the \( v \)th joint candidate located at \( p_v \) with label \( u_v \) and the \( w \)th joint candidate located at \( p_w \) with label \( u_w \) in terms of whether they are from the same person.

Given learnable parameters \( \Theta \), we aim to model the conditional probability \( p(p, u, b|I, \Theta) \) for solving the multi-person pose estimation problem. To this end, the proposed GPN adopts a generative model to produce person partitions implicitly and infers joint configuration for each person partition locally, as explained in Introduction. Hence, we introduce partition latent variables \( g = \{g_1, g_2, \ldots, g_M\} \) to encode the set of person partitions and thus the conditional probability \( p(p, u, b|I, \Theta) \) can be written as:

\[
p(p, u, b|I, \Theta) = \sum_g p(p, u, b, g|I, \Theta) = \sum_g p(p|I, \Theta)p(g|I, \Theta, p)p(u, b|I, \Theta, p, g),
\]  

(1)

where \( p(p|I, \Theta)p(g|I, \Theta, p) \) models the generative process of person partitions based on joint candidates. Our aim is to maximize the above conditional probability to find the optimal configurations of human poses for multiple persons in the image I. Instead of maximizing the probability marginalized w.r.t. all possible \( g \), we propose to maximize the lower bound by only considering the single “optimal” partition. Such approximation could reduce the complexity significantly without hurting the performance. Based on Eqn. (1), we can get:

\[
p(p, u, b|I, \Theta) \geq p(p|I, \Theta) \left\{ \max_g p(g|I, \Theta, p) \right\} p(u, b|I, \Theta, p, g).
\]  

(2)

Here, we find the optimal solution by maximizing the induced lower bound of \( p(p, u, b|I, \Theta) \), instead of maximizing the summation, i.e.,

\[
\tilde{p}, \tilde{u}, \tilde{b}, \tilde{g} = \arg \max_{p, u, b, g} p(p|I, \Theta)p(g|I, \Theta, p)p(u, b|I, \Theta, p, g).
\]  

(3)

The person partition \( g \) disentangles independent joints and reduces inference complexity—only the joints falling in the same partition have non-zero proximities. Then \( p(p, u, b, g|I, \Theta) \) can be further factorized as following:

\[
p(p, u, b, g|I, \Theta) = p(p, g|I, \Theta) \prod_{k \in g} p(u_k, I, \Theta, p, g_k)p(b_k, I, \Theta, p, g_k, u),
\]  

(4)

where \( u_k \) denotes the labels of all joints falling in the partition \( g_k \) and \( b_k \) denotes the corresponding proximities. Based on the above equations, we define \( p(p, u, b, g|I, \Theta) \) as a Gibbs distribution:

\[
p(p, u, b, g|I, \Theta) \propto \exp\{-E(p, u, b, g)\},
\]  

(5)

where \( E(p, u, b, g) \) is the energy function for the joint distribution \( p(p, u, b, g|I, \Theta) \). Its explicit form is derived from Eqn. (4) accordingly as:

\[
E(p, u, b, g) = -\varphi(p, g) - \sum_{k \in g} \left( \sum_{p_v \in g_k} \psi(p_v, u_v) + \sum_{p_v, p_w \in g_k} \phi(p_v, u_v, p_w, u_w) \right).
\]  

(6)

Here, \( \varphi(p, g) \) represents the score of person partition set \( g \) generated from \( p \), \( \psi(p_v, u_v) \) represents the score for labeling position \( p_v \) with \( u_v \), and \( \phi(p_v, u_v, p_w, u_w) \) represents the score for position \( p_v \) with label \( u_v \) and \( p_w \) with label \( u_w \) belonging to the same person, i.e., characterizing \( b_{u,v} \). In the following sections, we will illustrate the details for joint candidate detection, generative person partition, and algorithm to optimize the energy function.
2.2 Joint candidate detection

To reliably detect human body joints, we use the confidence maps to represent the confidences of joints present at each location in the image, which are constructed by modeling the joint locations as Gaussian peaks. We use $C_j$ to denote the confidence map for the $j$th joint with $C_j^i$ being the confidence map of the $j$th joint for the $i$th person. For a position $p_v$ in the given image, $C_j^i(p_v)$ is calculated as following:

$$C_j^i(p_v) = \exp \left\{ - \frac{\|p_v - p_j^i\|^2}{2\sigma^2} \right\}, \quad (7)$$

where $p_j^i$ denotes the position of the $j$th joint of the $i$th person, $\| \cdot \|_2$ is the Euclidean distance, and $\sigma$ is a constant which is empirically chosen to control the variance of Gaussian distribution. The target confidence map to be predicted is an aggregation of peaks of all people in a single map. Here, we choose to take the maximum of confidence maps rather than average to remain distinctions between close-by peaks, that is:

$$C_j(p_v) = \max_i C_j^i(p_v). \quad (8)$$

Previous works [17, 18] have shown that adding additional parts on limbs can help improve the detection accuracy for joints. Motivated by this, we also insert additional parts on the middle points of limbs as auxiliary guidance for joint detection. During testing time, we first find peaks with confidence score greater than a given threshold $\tau$ on predicted confidence maps $C$ for all types of joints. Then we perform non-maximum suppression to find joint candidate set $\hat{p} = \{p_1, p_2, \ldots, p_N\}$.

2.3 Generative person partition

In this section, we introduce the person partition generator based on Hough Transform to localize centroids of multiple persons through exploiting the cues from joint candidates. Hough transformation based approaches have been successfully used to the problem of part-based category-level object detection [19–21], due to its robustness to deformation, noise and various kinds of occlusion. This nature of Hough transform can address the difficulties facing by multi-person pose estimation problem. We adapt the Generalized Hough Transform framework for generating person partitions for multi-person pose estimation task, which generates person partitions and their composing joint candidates via employing joint detectors only once, overcoming the drawback of high joint detection complexity of top-down solution technique.

To achieve this goal, we construct the Hough space based on joint candidates which is parameterized by the person centroid, considering person centroid is stable and reliable to discriminate person instances even in presence of some extreme poses. This is different from traditional Hough spaces built on edge pixels, image patches, or image regions. We denote the constructed Hough space as $\mathcal{H}$, where each point $h_s \in \mathcal{H}$ corresponds to a hypothesis about the centroid of person instance. Joint candidates can vote for the centroid hypothesis of person since they are tightly related in the articulated kinematics sense. For instance, a head candidate might add votes for the presence of a person’s centroid to the location just below it. Joint candidates do not provide evidence for the exact centroid of person instance and their votes are distributed over many different centroid hypotheses in the Hough space. The probability of generating person partition $g_s$ at location $h_s$ is calculated by summing the votes from different joint candidates together, that is:

$$p(g_s | h_s) \propto \sum_j w_j \left( \sum_{p_v \in \hat{p}} \mathbb{1}[C_j(p_v) \geq \tau] \exp\{-\|f_j(p_v) - h_s\|^2_2\}\right), \quad (9)$$

where $\mathbb{1}[\cdot]$ is the indicator function, and $w_j$ is the weight of the $j$th joint. We set $w_j = 1$ for all joints assuming all kinds of joints equally contribute to the localization of person partitions due to the fact of unconstrained shapes of human body and uncertainties of presence of different joints. $f_j : \mathbb{R}^2 \rightarrow \mathcal{H}$ is the transformation function from joint location to Hough space. We formulate $f_j$ in a dense manner by constructing Hough transformation for every pixel in the image. We use $T_j^i$ to denote the transformation map for the $j$th joint of the $i$th person, which is defined as following:

$$T_j^i(p_v) = \begin{cases} o_{j,v}^i & \text{if } p_v \in \mathcal{N}_j^i, \\ 0 & \text{otherwise}, \end{cases} \quad (10)$$

$$o_{j,v}^i = \frac{(p_v^i - p_c) - (x_v^i - x_c, y_v^i - y_c)}{Z},$$
where $p_v^i$ denotes the center position of the $i$th person, $Z = \sqrt{H^2 + W^2}$ is the normalization factor, $H$ and $W$ are the height and width of image $I$, $N_v^i = \{p_v \mid \|p_v - p_v^i\|_2 \leq r\}$ denotes the neighbor positions of the $j$th joint of the $i$th person, and $r$ is a constant to define the neighborhood size. Then, we define the transformation map $T_j$ for the $j$th joint as the average for all persons by:

$$T_j(p_v) = \frac{1}{N_v} \sum_i T_j^i(p_v),$$

(11)

where $N_v$ is the number of non-zero vectors at position $p_v$ across all people. During testing time, after generating the predicted transformation map $T_j$, we define $f_j$ for position $p_v$ as:

$$f_j(p_v) = p_v + ZT_j(p_v).$$

(12)

After generating $p(g_i|h_i)$ for each point in the Hough space, we define the score $\varphi(p, g)$ as:

$$\varphi(p, g) = \sum_i \log p(g_i|h_i),$$

(13)

then the problem of person partition generation is converted to find peaks in the Hough space. Due to there are no priors about the number of people in the image, hence, to find peaks, we adopt the standard Agglomerative Clustering algorithm [20], which can automatically determine the number of clusters, to cluster the cast votes. We denote the vote set as $h = \{h_v \mid h_v = f_j(p_v), C_j(p_v) \geq \tau, p_v \in \tilde{p}\}$, and use $C = \{C_1, \ldots, C_M\}$ to denote the clustering result on $h$, where $C_i$ represents the $i$th cluster and $M$ is the number of clusters. We assume that the set of joint candidates casting votes in each cluster is corresponding to a person partition $g_i$, which is defined by:

$$g_i = \{p_v \mid p_v \in \tilde{p}, C_j(p_v) \geq \tau, f_j(p_v) \in C_i\}.$$

(14)

2.4 Local greedy inference

According to Eqn. (5), we maximize the conditional probability $p(p, u, b, g|I, \Theta)$ by minimizing the energy function $E(p, u, b, g)$ defined in Eqn. (6). We propose to optimize $E(p, u, b, g)$ in two sequential steps: first generating person partition set based on joint candidates; then, conducting inference for joint configurations of person instances in each person partition locally, which decouples the joint partition complexity and overcomes the drawback of bottom-up approaches. After generating each person partition according to Eqn. (14), the score $\varphi(p, g)$ becomes a constant. We use $g$ to denote the generated partition set, then the optimization problem can be simplified as:

$$\hat{u}, \hat{b} = \arg\min_{u, b} \left( -\sum_{g_i \in g} \sum_{p_v \in g_i} \psi(p_v, u_v) + \sum_{p_v, p_w \in g_i} \phi(p_v, u_v, p_w, u_w) \right).$$

(15)

We assume the pose estimation in each person partition is independent, then we conduct optimization for each person partition separately. In the following, we introduce our local greedy inference algorithm to the minimization problem defined in Eqn. (15) for multi-person pose estimation. Given a person partition $g_i$, we define the unary term $\psi(p_v, u_v)$ to be the confidence score at $p_v$ generated from the $u_v$th joint detector, that is:

$$\psi(p_v, u_v) = \tilde{C}_{u_v}(p_v).$$

(16)

The binary term $\phi(p_v, u_v, p_w, u_w)$ is defined by the similarity score of votes of two joint candidates in the Hough space, given by:

$$\phi(p_v, u_v, p_w, u_w) = 1[C_{u_v}(p_v) \geq \tau]1[C_{u_w}(p_w) \geq \tau] \exp{-\|h_v - h_w\|_2^2},$$

(17)

where $h_v = p_v + ZT_{u_v}(p_v)$ and $h_w = p_w + ZT_{u_w}(p_w)$. It is hard to directly optimize the minimization problem defined in Eqn. (15). Hence, for efficient inference, we adopt a greedy strategy for this task. The proposed greedy inference algorithm can guarantee the energy monotonically decreases and eventually converges to a lower bound. Specifically, we iterate through each joint one by one, first considering joints around the torso and gradually moving out to the limb. We first select neck as the start joint for inference. For a given neck candidate, we use its corresponding vote in the Hough space as the center of person instance. Then, we select the head top candidate closest to
Figure 2: Architecture of the designed Generative Partition Network, aiming to learn joint detector and Hough transformation simultaneously. We adopt multi-stage design to iteratively refine joint detection and recurrently correct errors of Hough transformation. We also add a feature pyramid branch to improve localization accuracy by combining high-resolution, semantically weak features with low-resolution, semantically strong features via a top-down pathway and lateral connections.

the person center and associate them to the same person. In addition, we will also update the center position of person instance by averaging the center hypotheses. We loop through all other types of joint candidates to associate with a person instance and update the center of person. After one loop, we will get a person instance. We follow the same strategy to find all person instances. After parsing person by neck as root, if there are still remaining candidates, then we move to torso as the root and inference person instance. After all candidates are associated with certain person, our algorithm will terminate. The overall local greedy inference algorithm is summarized in Algorithm 1.

3 Learning joint detector and Hough transformation with CNNs

The proposed GPN employs a Convolutional Neural Network architecture to jointly learn the joint detector and Hough transformation, which is shown in Figure 2. We design a multi-stage network with two separate branches—joint detection branch and Hough transformation branch—to learn the joint detector and Hough transformation simultaneously. In the first stage, we adopt VGG-19 to extract features. In every subsequent stage, we concatenate the predictions from two branches in last stage with image features extracted in first stage as input, providing contextual information to iteratively refine the confidence maps by learning implicit spatial relationships between different body joints and recurrently correct errors of Hough transformation maps by introducing top-down feedbacks. We add intermediate supervision at the end of each stage to overcome the gradient vanishing problem. We use L2 loss between the predictions and the targets as described for both branches, defined as following:

\[ L_t^1 = \sum_j \sum_v \| \tilde{C}_t^j(p) - C_t(p) \|_2^2, \]

\[ L_t^2 = \sum_j \sum_v \| \tilde{T}_t^j(p) - T_t(p) \|_2^2, \] (18)

where \( \tilde{C}_t^j \) and \( \tilde{T}_t^j \) represent the predicted confidence and Hough transformation maps at the 4th stage (1 \( \leq t \leq 6 \)), respectively. We also add a feature pyramid branch in our network architecture for combining high-resolution, semantically weak features with low-resolution, semantically strong features via a top-down pathway and lateral connections. Specifically, we pass the features from conv3_4 layer to the following layer, and an addition operation is adopted to combine with features for joint detection. The feature dimension is fixed as 256. We also pass the features to subsequent stages for guiding the estimation. Loss similar to previous joint detection is added in each stage as:

\[ L_t^3 = \sum_j \sum_v \| \tilde{C}_{t,2\times}^j(p) - C_{t,2\times}(p) \|_2^2, \] (19)

where \( C_{t,2\times} \) denotes the 2 times enlarged confidence map. The overall loss function is defined as:

\[ L = \sum_{t=1}^{T} (L_t^1 + \alpha L_t^2 + \beta L_t^3), \] (20)

where \( \alpha \) and \( \beta \) are weighting factors that is empirically found to work best when \( \alpha = \beta = 1 \).
Algorithm 1 The local greedy inference algorithm for multi-person pose estimation.

**Input:** joint candidates \( \tilde{p} \), person partitions \( \tilde{g} \), predictions \( \tilde{C} \) and \( \tilde{T} \), threshold \( \tau \).

**Initialization:** \( R \leftarrow \emptyset \)

for \( i = 1 \) to \( M \) do
  while \( g_i \neq \emptyset \) do
    \( P \leftarrow \emptyset \)
    for \( j = 1 \) to \( K \) do
      if \( \tilde{p}_j = \emptyset \) then
        \( \tilde{p}_j \leftarrow \arg \max_{p_v \in g_i} \tilde{C}_j(p_v) \)
      else
        \( \tilde{p}_j \leftarrow \arg \max_{p_v \in g_i} \mathbb{I}[\tilde{C}_j(p_v) \geq \tau] \exp\{-\|p_v + Z\tilde{T}_j(p_v) - c\|_2^2\} \)
      end if
      if \( \tilde{C}_j(\tilde{p}_j) \geq \tau \) then
        \( c \leftarrow \tilde{p}_j + Z\tilde{T}_j(\tilde{p}_j) \)
        \( P \leftarrow P \cup \{(\tilde{p}_j, j)\}, g_i \leftarrow g_i \setminus \{\tilde{p}_j\} \)
      end if
    end for
    \( R \leftarrow R \cup \{P\} \)
  end while
end for

**Output:** multi-person pose estimation results \( R \)

4 Experiments

**Datasets:** We evaluate our approach on two widely adopted benchmarks, MPII Multi-Person Pose dataset [22] and “We Are Family” (WAF) dataset [4]. The MPII Multi-Person Pose dataset consists of 3844 training and 1758 testing groups of multiple interacting individuals with 16 parts annotated for each person. In addition, it also contains more than 28,000 training samples for single person pose estimation. The WAF dataset consisting 525 internet images generated from the Google image search, where 350 images are used for training and 175 images for testing. Each person in the image of the WAF dataset is annotated with 6 line segments for the upper-body.

**Data augmentation:** We crop training samples on original images based on the person center. To prevent overfitting, we augment each training sample with rotation degrees in \([-40^\circ, 40^\circ]\), scaling with factors in \([0.7, 1.3]\), translational offset \([-40\text{px}, 40\text{px}]\), and horizontally mirror. We resize and pad training samples to the size \(368 \times 368\) pixels as input of CNNs.

**Experiment settings:** For training the model on the MPII Multi-Person Pose dataset, we randomly select 350 images excluded from the training set as validation set. We use the other training images from MPII Multi-Person Pose training set and all training samples for single person pose estimation for training our model. We train our model with Caffe [23] using the initial learning rate of \(4 \times 10^{-5}\). The parameters are optimized by RMSprop [24] algorithm. We train the model on the MPII dataset for 150 epochs. We then finetune the model of MPII on WAF for 30 epochs. During test time, we follow the standard routine to crop image patches with given position and group scale of test images. We also perform a scale search on 5 different scales of images to generate final results. We use “Mean Average Precision” (mAP) for evaluations on both the MPII and the WAF, suggested by [7].

4.1 Results

**MPII Multi-Person Pose dataset:** We show our evaluation results on MPII Multi-Person Pose dataset in Table 1 for the full testing set and the same subset of 288 testing images as in [6]. We can find that the proposed GPN achieves average 73.2% and 77.6% AP on full testing set and the subset, respectively, outperforming all published approaches. In particular, our approach achieves 13.7% AP
Table 1: Experimental results on the testing set of the MPII Multi-Person Pose dataset (AP).

| Method                  | Head   | Shoulder | Elbow   | Wrist  | Hip    | Knee    | Ankle   | Total  | Runtime [s] |
|-------------------------|--------|----------|---------|--------|--------|---------|---------|--------|-------------|
| Ours                    | 91.2   | 86.0     | 74.2    | 62.7   | 72.9   | 66.7    | 58.4    | 73.2   | 0.91        |
| Insafutdinov et al. [7] | 78.4   | 72.5     | 60.2    | 51.0   | 57.2   | 52.0    | 45.4    | 59.5   | 485         |
| Iqbal and Gall [1]      | 58.4   | 53.9     | 44.5    | 35.0   | 42.2   | 36.7    | 31.1    | 43.1   | 10          |
| Subset of 288 images as in [6] |
| Ours                    | 92.9   | 91.3     | 78.3    | 70.0   | 76.0   | 71.2    | 65.8    | 77.6   | 0.89        |
| Insafutdinov et al. [7] | 87.9   | 84.0     | 71.9    | 63.9   | 68.8   | 63.8    | 58.1    | 71.2   | 230         |
| Iqbal and Gall [1]      | 70.0   | 65.2     | 56.4    | 46.1   | 52.7   | 47.9    | 44.5    | 54.7   | 10          |
| Pishchulin et al. [6]   | 73.1   | 71.7     | 58.0    | 39.9   | 56.1   | 43.5    | 31.9    | 53.5   | 57995       |

improvement compared with the closest competitor on the full testing set, and 6.4% AP on the subset. Our approach also achieves the best performance for all joints. These results show the effectiveness of our approach. We also conduct analysis on the computational speed of our approach presented in Table 1. We can find that our approach achieves three orders magnitude faster than the state-of-the-art bottom-up approach [7]. This shows the efficiency of our approach. Qualitative results on MPII dataset are shown in Figure 3.

WAF dataset: We show our evaluation results on WAF dataset in Table 2. We can observe that our approach achieves superior performance over the state-of-the-arts. Our approach achieves overall 84.8% AP which achieves 3.4% improvement to the bottom-up method proposed in [7]. We can also see that our approach achieves the best performance for all upper-body joints, the largest improvement in performance is achieved on the elbow, about 10.3% higher than previous state-of-the-art accuracy. These results further verify the effectiveness of the proposed GPN to tackle the multi-person pose estimation problem.

Table 2: Experimental results on the testing set of the WAF dataset (AP).

| Method                  | Head   | Shoulder | Elbow   | Wrist  | Total  |
|-------------------------|--------|----------|---------|--------|--------|
| Ours                    | 93.1   | 82.9     | 83.5    | 79.9   | 84.8   |
| Insafutdinov et al. [7] | 92.6   | 81.1     | 75.7    | 78.8   | 82.0   |
| Pishchulin et al. [6]   | 76.6   | 80.8     | 73.7    | 73.6   | 76.2   |
| Chen and Yuile [25]     | 83.3   | 56.1     | 46.3    | 35.5   | 55.3   |

4.2 Ablation analyses

In this section, we conduct ablation analyses about the proposed approach on our MPII validation set, and the results are shown in Table 3.

GPN component-level analysis: We use “GPN-Full” to denote our full model, “GPN-w/o-Partition” to denote conducting inference on the whole image without person partitions, which is similar to bottom-up solutions, “GPN-w/o-LGI” to denote eliminating local greedy inference phase and partitioning joint candidates to person instances by finding the position with the maximal response for each type of joint in each person partition, which is similar to top-down solutions. We use “GPN-w/o-FPB” to denote generating joint candidates from joint detection branch rather than feature pyramid branch. From Table 3, “GPN-Full” achieves 76.7% AP on the validation set and the inference time for joint partition is 1.9ms. “GPN-w/o-Partition” achieves 76.4% AP with inference time 3.4ms. Comparing “GPN-w/o-Partition” with our full model, we can find local inference in each partition helps decouple the joint partition complexity, resulting in performance improvement and cost time decline. After eliminating local greedy inference phase as “GPN-w/o-LGI”, the performance drops to 74.5%, which shows the effectiveness of local greedy inference for estimating poses, by handling possible redundancies and false alarms of joint candidates based on affinity cues in the Hough space. “GPN-w/o-FPB” achieves 75.6% AP, comparing with our full model, which demonstrates the feature pyramid branch can help to achieve more accurate localization results for joint candidates.

1 The runtime time is measured on CPU Intel I7-5820K 3.3GHz and GPU TITAN X (Pascal).
Table 3: Results of ablation experiments on our MPII validation set (AP).

| Method        | Head | Shoulder | Elbow | Wrist | Hip  | Knee  | Ankle | Total | InferTime [ms] |
|---------------|------|----------|-------|-------|------|-------|-------|-------|----------------|
| GPN-Full      | 92.7 | 87.8     | 78.0  | 68.2  | 77.2 | 71.4  | 61.6  | 76.7  | 1.9            |
| GPN-w/o-Partition | 92.5 | 87.6     | 77.4  | 68.2  | 77.0 | 71.2  | 60.9  | 76.4  | 3.4            |
| GPN-w/o-LGI   | 88.3 | 85.0     | 76.4  | 66.6  | 74.8 | 69.8  | 60.3  | 74.5  | -              |
| GPN-w/o-FPB   | 91.7 | 87.4     | 76.1  | 66.2  | 76.8 | 70.7  | 60.2  | 75.6  | -              |
| GPN-3-Stage   | 92.0 | 87.9     | 77.3  | 65.9  | 76.9 | 70.3  | 58.8  | 75.6  | -              |
| GPN-2-Stage   | 91.4 | 87.3     | 76.6  | 65.2  | 76.8 | 69.9  | 58.0  | 75.0  | -              |
| GPN-1-Stage   | 88.7 | 85.7     | 74.0  | 61.8  | 74.3 | 63.7  | 52.5  | 71.5  | -              |

Multi-stage network analysis: To explore the effects of multi-stage network on learning Hough transformation, we conduct experiments as following: we fix the generation of joint candidates from the 6th stage of feature pyramid branch, and use the Hough transformation maps at different stages for producing estimation results, denoted as “GPN-\(t\)-Stage”, \(t \in \{1, 2, 3\}\) representing the stage index to generate Hough transformation maps. We can see that the performance increases monotonically as the increment of stages, and our final result with the 6th stage achieves about 7.3% improvement comparing with the first stage (76.7% vs 71.5%), showing the multi-stage design can refine Hough transformation maps by recurrently correcting their errors via top-down feedbacks.

Figure 3: Qualitative results on MPII dataset. Our approach provides accurate and robust multi-person pose estimation even in some challenging conditions, e.g., joint occlusions, people overlappings, pose changes, appearance variations, cluttered backgrounds.

5 Conclusion

In this paper we have presented the Generative Partition Network to address the multi-person pose estimation problem. It models the person partition as a generative process inferred from the votes of joint candidates in Hough space and performs local inference to estimate poses for person instances. The proposed approach has low joint detection and joint partition complexities, overcoming major drawbacks of pure top-down and bottom-up solutions. We also design a multi-stage convolutional neural network with feature pyramid branch to jointly learn the joint detector and the Hough transformation. The significant improvements in both performance and speed on two benchmarks multi-person pose evaluation demonstrate the effectiveness and efficiency of the proposed approach.
References

[1] U. Iqbal and J. Gall, “Multi-person pose estimation with local joint-to-person associations,” in ECCV, 2016.
[2] H. Fang, S. Xie, Y. Tai, and C. Lu, “RMPE: Regional multi-person pose estimation,” arXiv preprint arXiv:1612.00137, 2016.
[3] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, “Towards accurate multi-person pose estimation in the wild,” arXiv preprint arXiv:1701.01779, 2017.
[4] M. Eichner and V. Ferrari, “We are family: Joint pose estimation of multiple persons,” in ECCV, 2010.
[5] L. Pishchulin, A. Jain, M. Andriluka, T. Thomählen, and B. Schiele, “Articulated people detection and pose estimation: Reshaping the future,” in CVPR, 2012.
[6] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, “Deepcut: Joint subset partition and labeling for multi person pose estimation,” in CVPR, 2016.
[7] 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.
[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in CVPR, 2017.
[9] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, B. Andres, and B. Schiele, “Articulated multi-person tracking in the wild,” in CVPR, 2017.
[10] L. Ladicky, P. H. Torr, and A. Zisserman, “Human pose estimation using a joint pixel-wise and part-wise formulation,” in CVPR, 2013.
[11] E. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele, and B. Andres, “Joint graph decomposition & node labeling: Problem, algorithms, applications,” in CVPR, 2017.
[12] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, “Human pose estimation with iterative error feedback,” in CVPR, 2016.
[13] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in CVPR, 2016.
[14] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in CVPR, 2017.
[15] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in NIPS, 2015.
[16] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in ECCV, 2016.
[17] Y. Yang and D. Ramanan, “Articulated human detection with flexible mixtures of parts,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2878–2890, 2013.
[18] W. Yang, W. Ouyang, H. Li, and X. Wang, “End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation,” in CVPR, 2016.
[19] O. Barinova, V. Lempitsky, and P. Kholi, “On detection of multiple object instances using hough transforms,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 9, pp. 1773–1784, 2012.
[20] L. Bourdev and J. Malik, “Poselets: Body part detectors trained using 3d human pose annotations,” in ICCV, 2009.
[21] S. Maji and J. Malik, “Object detection using a max-margin hough transform,” in CVPR, 2009.
[22] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in CVPR, 2014.
[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding,” in Multimedia, 2014.
[24] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” COURSERA: Neural Networks for Machine Learning, 2012.
[25] X. Chen and A. L. Yuille, “Parsing occluded people by flexible compositions,” in CVPR, 2015.