2D human pose estimation is an important task in computer vision and machine learning, whose objective is to localize and assembly human skeletal keypoints for all human instances appeared in images. It is a challenging task due to the high structural and appearance variations of human poses and the complicated interactions between humans and objects (e.g. occlusions). Before the resurgence of deep neural networks (a.k.a. deep learning), human pose estimation is usually formulated under the structured prediction framework such as structural Support Vector Machine, which exploits pose relevant structural information in designing objective functions and optimization methods. In the deep learning era, human pose estimation is often cast as classification problem with two schema: The top-down scheme adopts a divide-and-conquer paradigm by first detecting humans (bounding box classification) in an image and then estimating the keypoints via keypoint classification for each detected human. The bottom-up scheme first detect all keypoints in an image (via keypoint classification) and then assembly them into different human skeletons using some greedy matching algorithm.

In this prospectus, 2D human pose estimation is addressed by integrating convolutional neural networks (ConvNets) and structural information in an end-to-end framework. The powerful discriminative features learned by ConvNets are exploited to explore the space of rich structural information of human poses. The proposed framework utilizes two key observations for keypoint prediction tasks from the traditional structured prediction formulation and the state-of-the-art deep learning formulation respectively: On the one hand, developing global pose consistent structural encoding of keypoints is crucial in improving the performance and interpretability of human pose estimation. On the other hand, maintaining high-resolution feature maps in the architectures of ConvNets is the key to leverage the expressive power of ConvNets for pixel-level keypoint based prediction tasks. The proposed framework studies the integration of structural information and high-resolution ConvNets in a comprehensive way. In this prospectus, limb guided attraction vector fields and multi-kernels are developed as structural information and integrated with a recently proposed high-resolution ConvNet, i.e., HRNets (27). Attention mechanism is also integrated in the proposed deep ConvNet architecture. The proposed methods have
been tested with state-of-the-art performance obtained for single 2D human pose estimation tasks in three widely used datasets: the FLIC (Frames Labeled In Cinema) dataset, the MPII human pose dataset and multi-person MS COCO dataset. In the on-going work, more structural information are investigated including coarse-to-fine and whole-part pose decompositions. Novel neural architectures for high-resolution ConvNets are also studied including deep AND-OR Grammar networks (52).
2D Human Pose Estimation by Integrating Convolutional Neural Networks and Structural Information

by
Liang Dong

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2019

APPROVED BY:

Christopher Healey
Edgar Lobaton

Tianfu Wu
Co-chair of Advisory Committee

Min Chi
Co-chair of Advisory Committee
# TABLE OF CONTENTS

## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2 LITERATURE REVIEW AND CONTRIBUTION SUMMARY</td>
<td>10</td>
</tr>
<tr>
<td>Chapter 3 Multi-Kernel Heads with Self-Attention</td>
<td>28</td>
</tr>
</tbody>
</table>

## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2 LITERATURE REVIEW AND CONTRIBUTION SUMMARY</td>
<td>10</td>
</tr>
<tr>
<td>Chapter 3 Multi-Kernel Heads with Self-Attention</td>
<td>28</td>
</tr>
</tbody>
</table>

## Chapter 1 INTRODUCTION

1.1 Problem Definition and Challenges | 1 |
1.2 Overview of Benchmarks | 3 |
  1.2.1 Frames Labeled In Cinema (FLIC) Benchmark | 4 |
  1.2.2 Leeds Sports Pose Benchmark | 4 |
  1.2.3 MPII Benchmark | 5 |
  1.2.4 Common Objects in Context Dataset(COCO) Benchmark | 6 |

## Chapter 2 LITERATURE REVIEW AND CONTRIBUTION SUMMARY

2.1 Overall Framework | 10 |
  2.1.1 Top-down Pipeline | 11 |
  2.1.2 Bottom-up Pipeline | 12 |
2.2 Traditional Method for Human Pose Estimation | 13 |
2.3 Methods Based on Deep Learning | 15 |
  2.3.1 DeepPose | 16 |
  2.3.2 Convolutional Pose Machines | 16 |
  2.3.3 Stacked Hourglass Network | 17 |
  2.3.4 Cascaded Pyramid Networks and Improvement | 18 |
  2.3.5 Simple Baselines | 19 |
  2.3.6 Deep High-Resolution Representation | 20 |
  2.3.7 Pose Refinement | 21 |
  2.3.8 Self-supervised | 23 |
  2.3.9 Attention Mechanism | 24 |
  2.3.10 The Development Of Backbone ConvNet | 24 |
2.4 The Proposed Work and Contributions | 26 |

## Chapter 3 Multi-Kernel Heads with Self-Attention

3.1 Motivation | 29 |
3.2 Method Overview | 29 |
  3.2.1 Multi-kernel Heads | 30 |
  3.2.2 Fusion | 32 |
  3.2.3 Self-Attention | 33 |
3.3 Experiments | 33 |
  3.3.1 Performance on COCO Val2017 Dataset | 34 |
  3.3.2 Visualization of multi-scale feature maps | 35 |
Chapter 4 Multi-Branch Heads with Limb Guided Attraction Field

4.1 Motivation

4.2 Structural Information for Limb

4.2.1 Part Affinity Field for Limb

4.2.2 Attraction Field Map for Limb

4.3 Method Overview

4.4 Experiments

4.4.1 Experiments on FLIC Dataset

4.4.2 Experiments on MPII Dataset

4.5 Summary

Chapter 5 Summary and Future Work

5.1 Summary

5.2 Future Work

References

BIBLIOGRAPHY

APPENDICES

Performance on FLIC dataset
| Table 2.1 | Comparision between traditional method and DeepConvNet | 27 |
| Table 3.1 | Performance analysis on MS COCO dataset on AP56 bounding box | 34 |
| Table 3.2 | Performance analysis on MS COCO dataset on groundtruth bounding box | 34 |
| Table 4.1 | FLIC results (PCK@0.2) | 54 |
| Table 4.2 | Performance of PCKh@0.1 on MPII validation dataset | 55 |
| Table 4.3 | Performance of PCKh@0.5 on MPII validation dataset | 56 |
LIST OF FIGURES

Figure 1.1 2D Human Pose Estimation. The left figure shows the COCO keypoint format, it has 17 full body keypoints, the right figure is the result of testing image by the pose estimation algorithm by Cao (59). ................................................................. 2

Figure 1.2 Challenging scenes of Human Pose Estimation (54). Left figures represent large variation, middle figures represent strong articulation, right figures represent occlusion. ............................................................. 3

Figure 1.3 FLIC Dataset. FLIC dataset is annotated with 9 upper-body joint locations. The figure is the result of FLIC test dataset by the pose estimation algorithm by Tompson (23). ............................................................... 4

Figure 1.4 Leeds Sports Dataset. The dataset is annotated with 14 full-body joints. The figure is the result of testing images by the pose estimation algorithm by Tompson (23). ............................................................. 4

Figure 1.5 MPII Dataset. The dataset is annotated with 16 full-body joint locations. The figure is the result of testing images by the pose estimation algorithm by Wang (58). ................................................................. 6

Figure 1.6 MS COCO dataset (44). The Dataset is annotated with 17 full human body joints. The figure shows the annotated bounding box, keypoints and the links between keypoints. ............................................................. 7

Figure 1.7 Keypoint Similarity (31). The keypoint similarity of left eye is 0.25, the keypoint similarity of right ankle is 0.75. ................................................................. 8

Figure 2.1 Two Types of Pipelines for Human Pose Estimation. For 2D human pose estimation, the image is firstly generated to feature, then there are two branches: top-down pipeline and bottom-up pipeline. Top-down pipeline firstly does human detection and then single person pose estimation. Bottom-up pipeline does keypoint detection and then associate keypoints. .............. 11

Figure 2.2 Top-down Pipeline for Human Pose Estimation. The top-down pipeline firstly predicts bounding boxes of humans, then crops the human and feeds it to the single person pose estimation network. There are two main parts: human detection and single pose estimation. ......................... 12

Figure 2.3 Bottom-up Approach for Human Pose Estimation. The bottom-up pipeline jointly detects human keypoints and associates them to individual people. There are two main parts: keypoint association and keypoint detection. 13

Figure 2.4 Pictorial Structure (30) ........................................................................ 14

Figure 2.5 Articulated Pose Estimation with Flexible Mixtures-of-parts (55) .... 15
Figure 2.6  Deep Pose (3). The left figure is schematic view of the DNN-based pose regression. They visualize the network layers with their corresponding dimensions, where convolutional layers are in blue, while fully connected ones are in green. They do not show the parameter free layers. The right figure at stage \( s \) is a refining regressor applied on a sub image to refine a prediction from the previous stage.

Figure 2.7  Architecture and receptive fields of Convolutional Pose Machines (42). They show a convolutional architecture and receptive fields across layers for a CPM with any \( T \) stages. The pose machine (42) is shown in insets (a) and (b), and the corresponding convolutional networks are shown in insets (c) and (d). In (b) and (d) are repeated for all subsequent stages (2 to \( T \)). Below in inset (e) they show the effective receptive field on an image (centered at left knee) of the architecture, where the large receptive field enables the model to capture long-range spatial dependencies such as those between head and knees. (Best viewed in color.)

Figure 2.8  Stacked Hourglass Network (1). The network consists of eight stacked hourglass modules to do repeated bottom-up, top-down fusion.

Figure 2.9  Cascaded Pyramid Networks (56). The GlobalNet is based on the ResNet backbone. A RefineNet to explicitly address the hard keypoints. L2 loss* means L2 loss with online hard keypoints mining.

Figure 2.10  Simple Baselines (5).

Figure 2.11  High-Resolution Network (27).

Figure 2.12  PoseFix (14).

Figure 2.13  Data Distillation (18). In data distillation, ensembled predictions from a single model applied to multiple transformations of an unlabeled image are used as automatically annotated data for training a student model.

Figure 2.14  Multi context attention mechanism (51). (a) is the original heatmap and predicted pose. (b) is the attention map. (c) is the refined pose by attention, the part attention model could fix the double counting problem.

Figure 3.1  Single Kernel Head. The image firstly goes through ConvNet Feature Backbone such as SimpleBaseline (5) or HRNet (27), then map to the feature map with the number of the channels is the number of keypoints. The kernel size of the mapping is 3*3.

Figure 3.2  Motivation of Multi-Kernel Heads Design. Compared to Single Kernel head, the multi-kernel heads add a filter bank with multiple kernel sizes to generate heatmap bank. Then the final heatmap is generated by the fusion of the heatmap bank.
Figure 3.3 Method Overview of Multi-Kernel Heads. The filter bank of the multi-kernel heads support 9 kernels. The widths are 3, 5 and 7 and the heights are 3, 5 and 7.

Figure 3.4 The plots of different human body keypoints combination. The middle point is in the origin. The blue line represents the left keypoint, the red line represents the right keypoint. six combinations are listed in the figure.

Figure 3.5 Fusion of Multi-kernel Heads. Firstly self-attention mechanism works in the heatmap bank to recalibrate the relative importance of each heatmap, then the mean operation is adapted to get the final heatmap.

Figure 3.6 Self-attention Mechanism. The importance weight is calculated by adding global pooling and two fully connected layers to the heatmap bank. Then each heatmap in the heatmap bank is multiplied the weight.

Figure 3.7 Pose Estimation Result on COCO Validation Dataset. The orange rectangles represent the ground-truth human bounding box. The red color keypoints represent the keypoints in the left side while the green color keypoints represent the keypoints in the right side.

Figure 3.8 Visualization of multi-scale feature maps. Left figure is the test image from COCO 2017val dataset. Right figure is the pose estimation result with the kernel size of maximal response.

Figure 3.9 Visualization of Feature Maps in Feature Bank. The feature map generated by 9 different kernel sizes is shown in the figure. The size of the rectangle represents the size of the kernel to generate feature map.
Figure 3.10 Performance and Error Sensitivity to Occlusion and Crowding for HRNet (27). The PR curves showing the performance of obtained by progressively correcting errors of each type at the OKS evaluation threshold of .75 on the twelve Occlusion and Crowding Benchmarks. 38

Figure 3.11 Performance and Error Sensitivity to Occlusion and Crowding for Multi-kernel heads with self-attention mechanism. The PR curves showing the performance of obtained by progressively correcting errors of each type at the OKS evaluation threshold of .75 on the twelve Occlusion and Crowding Benchmarks. 39

Figure 3.12 Taxonomy of Keypoint Localization Errors. Jitter: left hip in (a); Inversion: inverted skeleton in (b), right wrist in (c); Swap: right wrist in (d), right elbow in (e); Miss: both ankles in (f). 40

Figure 3.13 Localization error evaluation. Good means that the object similarity between predicted keypoints and groundtruth keypoints is less than 0.5. The proposed multi-head heads with self-attention mechanism get better performance. 41

Figure 3.14 Inversion localization error comparison. 42

Figure 3.15 Jitter localization error comparison. 43

Figure 3.16 Miss localization error comparison. 44

Figure 3.17 Swap localization error comparison. 45

Figure 4.1 Motivation for Multi-stage Multi-branch Heads. In the top-down pipeline, the proposed heads encode structural information such as local structural information and global structural information besides heatmap branch. 47

Figure 4.2 Pixel Assignment Method. Every pixel in the entire image is attracted to one and only one limb based on the distance. 50

Figure 4.3 Attraction Field Map (35). The left image is the attraction field map for $d_x$, the right image is the attraction field map for $d_y$. 50

Figure 4.4 Line segment detection for 2D Human Pose. The left image is from FLIC test dataset. The right image is the fitted line segment in testing image. 51

Figure 4.5 Proposed Multi-stage Multi-Branch Heads. The proposed heads have three stages and three branches: confidence heat maps, part affinity fields for local structural information and attraction field maps for global structural information. 52

Figure 4.6 Pose Estimation Result on FLIC test dataset. The green circle represents ground truth keypoints while the red plot represents the predicted keypoints. 54

Figure 4.7 Pose Estimation Result on MPII validation dataset. The blue rectangles represent ground truth while the red represents the predicted keypoint. 55

Figure 4.8 Performance Plot on MPII validation dataset for the proposed method. The are seven parts listed in the figure. The performance is worse in hard keypoints such as hip and ankle. 56
CHAPTER

1

INTRODUCTION

2D human pose estimation is an important and hot topic in computer vision. There are many applications based on 2D human pose estimation. Firstly, 2D pose estimation is important intermediate information for human action recognition. Secondly, It could be used for human computer interaction in virtual reality or augmented reality. In the last, it could supply more information for 3d human pose estimation. For example, 3D human pose estimation could be implemented by 2D Pose Estimation and Matching. However, it is a challenging problem. There are many challenging scenes like large variation of human pose, human shape, and background. There may be strong articulation, self occlusion or other occlusions.

1.1 Problem Definition and Challenges

The objective of 2D human pose estimation algorithm is to localize and assemble human skeletal keypoints (e.g., ankle, wrist, etc.) for all human instances appearing in images, as shown in figure 1.1.
2D human pose estimation has wide applications such as human action recognition, human-computer interaction, animation, auto-driving monitoring, etc. It is a very challenging problem: an unknown number of people could appear at different position or scale, there could be complex spatial interference, or individuals may have self-occlusion and strong articulations. Some examples of challenging scenes for human pose estimation is shown in figure 1.2.
The human pose estimation problem could be divided into two categories: single person pose estimation (there is only one person in the image) and multi-person pose estimation.

There are mainly two approaches to solve the problem of single-person human pose estimation: classical probability graphical model and deep ConvNet.

1.2 Overview of Benchmarks

2D human pose estimation is formulated under the supervised learning framework. In order to compare different algorithms, researchers have created several public datasets and defined the evaluation metrics. Benchmarks (datasets and evaluation metric) are the main driving force in promoting computational methods. There are four popular evaluation metrics for human pose estimation: Percentage of Correct Parts (PCP), Standard Percentage of Correct Keypoints (PCK), Head-Normalized Probability of Correct Keypoint (PCKh) and
Object Keypoint Similarity (OKS). In order to do a fair comparison, several annotated public datasets for human pose estimation are created. There are four popular datasets widely used to evaluate the performance of human pose estimation algorithm. Following are the brief introduction of the four benchmarks.

### 1.2.1 Frames Labeled In Cinema (FLIC) Benchmark

The FLIC dataset is composed of 5003 still RGB images with 2D ground-truth joint information generated using Amazon Mechanical Turk. The dataset contains 3987 for training, 1016 for testing. Some samples from FLIC dataset are shown in [figure 1.3](image). Each image has been annotated with 9 joint locations. The FLIC dataset contains many frames with more than a single person, while the joint locations from only one person in the scene are labeled. Researchers usually used PCK (standard Percentage of Correct Keypoints metric which reports the percentage of detections that fall within a normalized distance of the ground truth) to evaluate the performance of different algorithms in FLIC dataset. Stacked hourglass network (1) achieves 99.0 in elbow and 97.0 in wrist. The best performance is 3-level deeply learned compositional models (11) with 99.5 in elbow and 98.5 in wrist.

![Figure 1.3: FLIC Dataset.](image)

The figure is the result of FLIC test dataset by the pose estimation algorithm by Tompson (23).

### 1.2.2 Leeds Sports Pose Benchmark

The Leeds Sports Pose (LSP) dataset contains 2,000 pose annotated images of mostly sports people gathered from Flickr, 1000 for training and 1000 for testing. Some samples from LSP dataset are shown in [figure 1.3](image). The images have been scaled such that the most
prominent person is roughly 150 pixels in length. Each image has been annotated with 14 joint locations.

![Figure 1.4: Leeds Sports Dataset. The dataset is annotated with 14 full-body joints. The figure is the result of testing images by the pose estimation algorithm by Tompson (23).](image)

The LSP dataset has two evaluation settings, person-centric (PC) and observer-centric (OC). The state of art performance is 93.9% at PCK-0.2 by Wei Yang etc (47).

### 1.2.3 MPII Benchmark

The MPII Human Pose dataset comes out from Max Planck Institute for Informatics. It consists of images taken from a wide-range of real-world activities with full-body pose annotations. Each image has been annotated with 16 joint locations. There are around 25,000 images with 40,000 subjects, where there are around 2,000 images for validation and the remaining subjects for the training set. The dataset consists of images with 16 keypoints with full-body pose annotations. The images are taken from real-world activities. There are around 25,000 images with 40,000 subjects, where there are around 2,000 images for validation and the remaining subjects for training. The image size is cropped to 256*256 for fair comparison with other algorithm. The standard evaluation metric is the PCKh(head-normalized probability of correct keypoint) score. The state of art performance on test dataset is 92.3% of PCKh-0.5 achieved by high resolution representation (27) and deeply learned compositional models (46). Hong Zhang (16) presented two conceptually simple and yet computational efficient modules, namely Cascade Prediction Fusion and Pose Graph Neural Network, to exploit underlying contextual information. It achieved 92.5% of PCKh-0.5. Before that, Wei et at. (47) achieved 92% of PCKh-0.5 by a Pyramid Residual
Module to enhance the invariance in scales of deep convolutional neural networks. One picture from MPII dataset are shown in figure 1.5. There are multiple annotated persons in the image with 16 human body keypoints.

![Figure 1.5: MPII Dataset. The dataset is annotated with 16 full-body joint locations. The figure is the result of testing images by the pose estimation algorithm by Wang (58).](image)

The standard evaluation metric is the PCKh(head-normalized probability of correct keypoint) score. A joint is correct if it falls within $\alpha l$ pixels of the groundtruth position, where $\alpha$ is a constant and $l$ is the head size that corresponds to 60% of the diagonal length of the ground-truth head bounding box. PCKh@0.5 is usually used and PCKh@0.5 is when the threshold = 50% of the head bone link. In the proposed method, PCKh@0.1 is also reported to do detailed comparison.

### 1.2.4 Common Objects in Context Dataset(COCO) Benchmark

The MS COCO Keypoint Challenge (44) requires localization of multi-person keypoints in challenging uncontrolled conditions. The dataset contains over 200,000 images and 250,000 person instances labeled with 17 keypoints. The MS COCO train2017 dataset includes 57,000 images and 150,000 person instances. The val2017 set and test-dev2017 set contain 5000 images and 20,000 images, some samples are shown in figure 1.6.
The MS COCO dataset (44). The Dataset is annotated with 17 full human body joints. The figure shows the annotated bounding box, keypoints and the links between keypoints.

For MS COCO dataset, researchers usually use the OKS (object keypoint similarity) as the evaluation metric, which will be discussed in the following section. Given specific OKS, Average precision and average recall are calculated. The object keypoint similarity (OKS) is defined in the following equation. It is based on keypoint similarity.

\[
OKS = \frac{\sum_i \exp(-d_i^2/2s_i^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}
\]  

(1.1)

The \(d_i\) are the Euclidean distances between each corresponding ground truth and detected keypoint and the \(v_i\) are the visibility flags of the ground truth. To compute OKS, \(d_i\) is passed through an unnormalized Guassian with standard deviation \(s_i\), and \(s_i\) is the object scale and \(k_i\) is a per-keypoint constant and it controls falloff. Keypoint similarity of each keypoint ranges from 0 to 1. A demo of keypoint similarity is shown in figure 1.7.
These similarities are averaged over all visible keypoints. Predicted keypoints that are not labeled ($v_i=0$) do not affect the OKS. Accurate predictions will have $OKS = 1$ and predictions for which all keypoints are off by more than a few standard deviations $s_i$ will have $OKS = 0$. The OKS is analogous to the intersection over union. Given the OKS, average precision and average recall can be computed. Researchers usually report mean average precision and recall scores. The AP is the mean of average precision scores at 10 positions (OKS = 0.50, 0.55, . . . , 0.90, 0.95). $AP_{OKS=0.50}$ is the average precision at $OKS = 0.5$. $AP_{OKS=0.75}$ is the average precision at $OKS = 0.75$. $AP_{medium}$ is the mean of average precision scores for medium objects. $AP_{large}$ is the mean of average precision scores for medium objects. $AR$ is the mean of average large scores at 10 positions (OKS=0.50, 0.55, . . . , 0.90, 0.95). The next section discusses the developed algorithms to solve the 2D human pose estimation problem. The algorithms are evaluated on MS COCO dataset using $AP$, $AP_{OKS=0.50}$, $AP_{OKS=0.75}$, $AP_{medium}$, $AP_{large}$, $AR$ based on OKS.
For HRNet (27), the HRNet-W32 achieved 74.9 mAP. HRNet-W48 with extra data achieved 77 mAP. The best result of simple baselines model (5) achieves the state-of-the-art at mAP of 73.7 on COCO testdev split. Wenbo Li et al. proposed enhanced multi-stage pose networks (48) and achieved 76.1 mAP by three methods: (a) adopting the existing good network structure for down sampling path and a simple up sampling path, (b) aggregating features across different stages to strengthen the information flow and mitigate the difficulty in training, (c) adopting coarse-to-fine supervision strategy. The multi-stage pose network using an ensemble model and external data achieved 78.1 mAP.
Chapter 1 introduces 2D human pose estimation, the challenging scenes and its application. It discusses both classical approach and deep ConvNet approach for single-person pose estimation. Multi-person pose estimation including top-down and bottom-up methods is also introduced. This chapter focuses on the deep ConvNet architecture design for single-person pose estimation. This chapter also gives a literature review for milestones in deep ConvNet architecture design for single-person pose estimation and brief introduction of the development of backbone ConvNet in section one, which could be used in multi-person pose estimation.

2.1 Overall Framework

There are two types of pipelines to solve the 2D human pose estimation problem: top-down pipeline and bottom-up pipeline. The top-down pipeline firstly gets the human detection
result and does single person pose estimation in the bounding box. The other bottom-up pipeline gets all the keypoints first and then does keypoint association. The two types of pipelines are shown in figure 2.1.

2.1.1 Top-down Pipeline

The top-down pipeline firstly gets human bounding box and then regress human keypoints in the bounding box. These top-down approaches could directly leverage existing techniques for single-person pose like Stacked Hourglass Network (1), SimpleBaseline (5), High Resolution Representation Net (27). The top-down pipeline firstly predicts bounding boxes of humans and then crops the human and feeds it to the single person pose estimation network. Kaiming He proposed Mask-RCNN (24) to crop human feature map by ROIAlign and do keypoint detection based on the feature map. Xiao (5) combined the upsampling and convolutional parameters into deconvolutional layers in a much simpler way, without using skip layer connections. The top-down approach highly depends on human detection results, and it fails in crowded scenes. Also, the run time of these top-down approaches is proportional to the number of people: for each detection, a single-person pose estimator is run, and the more people there are, the greater the computational cost. The top-down architecture is shown in figure 2.2. The main concern is how to design the single person pose estimation network architecture.
Figure 2.2: Top-down Pipeline for Human Pose Estimation. The top-down pipeline firstly predicts bounding boxes of humans, then crops the human and feeds it to the single person pose estimation network. There are two main parts: human detection and single pose estimation.

2.1.2 Bottom-up Pipeline

The bottom-up pipeline jointly labels part detection candidates and associates them to individual people. The bottom-up approach architecture is shown in Figure 2.3. There are two branches: parts detection and parts association. The following is brief introduction of related work in the bottom-up approach. Deepcut (28) labeled part detection candidates and associated them to individual people by solving the integer linear programming problem. DeeperCut (9) improves the DeepCut (28) by introducing image-conditioned pair-wise terms. Cao etc. (59) presented the first bottom-up representation of association scores via Part Affinity Fields (PAFs), a set of 2D vector fields that encode the location and orientation of limbs over the image domain. The proposed method achieves realtime multi-person 2D pose estimation. Newell et al. (2) proposed associative embeddings which can be thought as tags representing each keypoints’ group. They grouped keypoints with similar tags into individual people. George Papandreou (12) added their network a separate pairwise mid-range 2-D offset field output to connect pairs of keypoints. They computed 2(K-1) such offset fields, one for each directed edge connecting pairs (k, l) of keypoints which are adjacent.
to each other in a tree-structured kinematic graph of the person. Nie et al. (53) proposed to partition all keypoint detections using dense regressions from keypoint candidates to centroids of persons in the image. Kocabas et al. (34) proposed a multi-task model and assigned joints to detected persons by a pose residual network. The bottom-up method is relatively vulnerable because they only consider small local regions and output smaller response heatmaps, but their speed is much higher than top-down method.

Figure 2.3: Bottom-up Approach for Human Pose Estimation. The bottom-up pipeline jointly detects human keypoints and associates them to individual people. There are two main parts: keypoint association and keypoint detection.

There are two approaches for single-person pose estimation: traditional approach and deep ConvNet. Traditional approach is mainly dominated by probabilistic graphical model. Methods based on deep learning forms the representation of human keypoints and designs deep convolutional neural network architecture for human pose recognition.

### 2.2 Traditional Method for Human Pose Estimation

Before deep ConvNet, human pose recognition is mainly dominated by hand-designed features and probabilistic graphical models. Traditional approaches for human pose esti-
mation are heavily based on the idea of part-based models, as pioneered by the pictorial structures model (30) of Fischler and Elschlager. The basic idea of pictorial structure is to represent an object by a collection of "parts" arranged in a deformable configuration (not rigid), shown in figure 2.4. The part-based model can model articulations well. The use of structure information for Human Pose Estimation comes through minimizing the error between an instance of the model and the given image. This is however achieved at the cost of limited expressiveness and does not take in global context into account. Yang and Ramanan (55) use a mixture model of parts which expresses complex joint relationships. Deformable part models are a collection of templates arranged in a deformable configuration and each model has a global template and part templates. While global optimality of the probabilistic graphical models is attractive, its ability to represent complex relations among parts and the expressive power of hand-crafted appearance feature are limited compared to deep ConvNet. 2D human pose estimation could also be classified as generative or discriminative method. Generative approaches aim at modeling all the intricacies of the image formation process, while discriminative models learn a direct mapping from image features to the pose space using training data (38).

Figure 2.4: Pictorial Structure (30)
2.3 Methods Based on Deep Learning

After deep ConvNet achieved big success in ImageNet, Toshev et al. (3) used the AlexNet architecture to directly regress spatial joint coordinates. Tompson et al. (23) learned pose structure by combining deep features along with graphical models. Carreira et al. (22) proposed the Iterative Error Feedback method to train ConvNet where the input is repeatedly fed to the network along with current predictions in order to refine the predictions. Newell et al. (1) stacked hourglass blocks to obtain an iterative refinement process and showed its effectiveness on single person pose estimation. Lifshitz et al. (20) used a probabilistic keypoint voting scheme from image locations to obtain agreement maps for each body part. Chu et al. (51) enhanced the stacked hourglass network (1) by integrating it with a multi-context attention mechanism. Ke et al. (29) proposed a multi-scale structure-aware network for single human pose estimation.

There are two categories of deep ConvNet design. Recently the best performing method is based on a single-stage backbone network. Cascaded pyramid network (56) is based on Res-Inception (6). The simple baseline approach (5) uses ResNet (25) as backbone, as pose estimation requires a high spatial resolution, upsampling or deconvolution is usually appended after the backbone networks to increase the spatial resolution of deep features. Another category of pose estimation methods adopts a multi-stage architecture. Each stage is a simple light-weight network and contains its own down sampling and up sampling paths. The feature maps between the stages remain a high resolution. All the stages are usually supervised simultaneously to facilitate a coarse-to-fine, end-to-end training. Representative
works include stacked hourglass network (1) and cascaded pyramid networks (56).

2.3.1 DeepPose

Alexander etc. (3) formulates the pose estimation as a DNN-based regression problem towards body joints. They present a cascade of a 7-layered generic deep ConvNet. The DeepPose architecture is shown in figure 2.6. The last stage has two fully connected layers to regress the human keypoints. It is the first human pose estimation algorithm based on deep ConvNet and it achieves state-of-art results on standard benchmarks like FLIC, LSP, and MPII datasets.

![Deep Pose](image)

Figure 2.6: Deep Pose (3). The left figure is schematic view of the DNN-based pose regression. They visualize the network layers with their corresponding dimensions, where convolutional layers are in blue, while fully connected ones are in green. They do not show the parameter free layers. The right figure at stage s is a refining regressor applied on a sub image to refine a prediction from the previous stage.

2.3.2 Convolutional Pose Machines

Convolutional pose machines (42) use heatmap as the representation of keypoint. It shows a systematic design for how convolutional networks can be incorporated into the pose machine framework for learning image features and image-dependent spatial models. It supplies intermediate supervision to suppress vanishing gradients problems. The convolutional pose machines architecture is shown in figure 2.7.
Figure 2.7: Architecture and receptive fields of Convolutional Pose Machines (42). They show a convolutional architecture and receptive fields across layers for a CPM with any $T$ stages. The pose machine (42) is shown in insets (a) and (b), and the corresponding convolutional networks are shown in insets (c) and (d). In (b) and (d) are repeated for all subsequent stages (2 to $T$). Below in inset (e) they show the effective receptive field on an image (centered at left knee) of the architecture, where the large receptive field enables the model to capture long-range spatial dependencies such as those between head and knees. (Best viewed in color.)

2.3.3 Stacked Hourglass Network

The design of the Stacked Hourglass is shown in figure 2.8, it is motivated by the need to capture information at every scale. While local evidence is essential for identifying features like faces and hands, a final pose estimate requires a coherent understanding of the full body. The Hourglass has repeated iterative u-shape structure. Firstly, convolutional and max pooling layers are used to process features down to a very low resolution. Then at each max pooling step, the network branches off and applies more convolutions at the original pre-pooled resolution. In the last step, the network begins the top-down sequence of upsampling and combination of features across scales after reaching the lowest resolution. To bring together information across two adjacent resolutions, Stacked Hourglass network (1) follow the process described by Tompson et al. It performs nearest neighbor upsampling of the lower resolution followed by an element-wise addition of the two sets of features.
The topology of the hourglass is symmetric, so for every layer present on the way down there is a corresponding layer going up. After reaching the output resolution of the network, two consecutive rounds of 1x1 convolutions are applied to produce the final network predictions. The output of the network is a set of heatmaps where for a given heatmap the network predicts the probability of a keypoint presence at each and every pixel.

Figure 2.8: Stacked Hourglass Network (1). The network consists of eight stacked hourglass modules to do repeated bottom-up, top-down fusion.

2.3.4 Cascaded Pyramid Networks and Improvement

Cascaded Pyramid Network (56) is the leading method on MS COCO 2017 keypoint challenge. It presents a novel network structure called Cascaded Pyramid Network (CPN) which targets to recognize hard keypoints like wrist, ankle in a lot of challenging cases, such as occluded keypoints, invisible keypoints and complex background, which cannot be easily well addressed. The network architecture is shown in figure 2.9. It involves skip layer feature concatenation and an online hard keypoint mining step. The cascaded pyramid network (CPN) integrates global pyramid network (GlobalNet) and pyramid refined network based on online hard keypoints mining (RefineNet). The GlobalNet learns a good feature representation based on feature pyramid network while the interesting part is that the RefineNet explicitly address the hard joints based on an online hard keypoints mining loss. The same team makes improvement of Cascaded Pyramid Network (56), which will be discussed in the following section.
Rethinking on Multi-Stage Networks for Human Pose Estimation (15) proposes several improvements, including the single-stage module design, cross stage feature aggregation, and coarse-to-fine supervision. The multi-stage network architecture is based on cascaded pyramid network. Although single-stage module design achieves better performance, it concludes that the current unsatisfactory performance of multi-stage networks for human pose estimation is due to the insufficiency in various network architecture design choices. The paper proposes three improvements and achieves state-of-art results.

- A good architecture GlobalNet from Cascaded Pyramid Networks (56) is adapted.
- A feature aggregation strategy is proposed to propagate information from early stages to the later ones.
- The usage of coarse-to-fine supervision is introduced. It adopts finer supervision in localization accuracy in later stages.

### 2.3.5 Simple Baselines

Simple Baselines (5) differs from Stacked Hourglass Network (1) and cascaded pyramid network (56) in how the high resolution feature maps are generated. Both works (Stacked Hourglass Network (1) and Cascaded Pyramid Network (56)) use upsampling to increase the feature map resolution and put convolutional parameters in other blocks. In contrary,
Simple BaseLines (5) combines the upsampling and convolutional parameters into deconvolutional layers in a much simpler way, without using skip layer connections. The commonality of the three methods is that three upsampling steps and also three levels of non-linearity (from the deepest feature) are used to obtain high resolution feature maps and heatmaps. In the last, the paper concludes that the crucial point is to obtain high resolution feature maps.

![Simple Baselines (5)](image)

**Figure 2.10:** Simple Baselines (5)

### 2.3.6 Deep High-Resolution Representation

After Simple Baseline (5), Ke Sun etc. adapts a new strategy and proposes high-resolution representation network (27). High-resolution Representation Network (27) is able to maintains high-resolution representations through the whole process. It avoids high-to-low process, multi-scale fusion and intermediate supervision in previous human pose estimation network design. The architecture is shown in figure 2.6. The architecture has two benefits in comparison to existing widely-used networks. The approach is able to maintain the high resolution instead of recovering the resolution through a low-to-high process as it
has high-to-low resolution subnetworks in parallel rather than in series as done in most existing solutions. Most existing fusion schemes aggregate low-level and high level representations. Instead, the High Resolution Representation (27) perform repeated multi scale fusions to boost the high-resolution representations with the help of the low-resolution representations of the same depth and similar level, and vice versa, resulting in that high-resolution representations are also rich for pose estimation. Consequently, their predicted heatmap is potentially more accurate.

Figure 2.11: High-Resolution Network (27)

2.3.7 Pose Refinement

There are many methods attempted to refine the estimated human keypoint result by introducing more structural information. Zhe proposed part affinity fields (PAFs) (59) to learn to associate body parts with individuals in multi-person human pose recognition. Carreria et al. (22) iteratively estimated error feedback from a shared weight model. The output error feedback of the previous iteration is then transformed into the input pose of
the next iteration, which is repeated several times for progressive pose refinement. Mihai Fieraru et al. introduced a pose refinement network (33) which takes as input both the image and a given pose estimate and learns to directly predict a refined pose by jointly reasoning about the input-output space. To model challenging cases, they also proposed a novel data augmentation procedure that allows to synthesize possible input poses and make the network learn to identify the erroneous body joint predictions and to refine them. Chu proposed (50) a CRF-CNN framework which can simultaneously model structural information in both output and hidden feature layers in a probabilistic way, and it is applied to human pose estimation. Chu also proposed (45) a structured feature learning framework to reason the correlations among body joints at the feature level in human pose estimation. The relationships between feature maps of joints are captured with the introduced geometrical transform kernels, which can be easily implemented with a convolution layer. Recently, Gyeongsik Moon (14) proposed a human pose refinement network that estimates a refined pose from a tuple of an input image and input pose. They did not adapt the pose refinement that was performed mainly through an end-to-end trainable multi-stage architecture in previous methods. Instead, they proposed a model-agnostic pose refinement method. The paper make a assumption that state-of-the-art 2D human pose estimation methods had similar error distributions and validate the assumption. The innovation of the paper is to use this error statistics as prior information to generate synthetic poses and use the synthesized poses to train the proposed pose refinement model. It is based on error statistics from empirical analysis. It does not need information or code of other human pose estimation methods. The input of Posefix (14) is the image and pose estimation result of the pose detection method. A demo of refined pose is shown in figure 2.12.
2.3.8 Self-supervised

Self-supervised learning is a subset of unsupervised learning. It can generate output labels from data objects by different views of the object or exposing a relation between parts of the object. Due to that, there is lately an increased interest to learn deep ConvNet based representations in an unsupervised manner that avoids manual annotation of visual data. There are a lot of data augmentation methods for images: flip, rotation, crop, scaling, affine transformation, perspective transformation and the sampling synthetic strategies such as SMOTE (36), Sampling Pairing (19), Mixup (17). Recently there are much data augmentation methods brought out for unsupervised learning and semi-supervised learning( (17), (8), (39), (43), (7))brought out for training deep ConvNet models.

Data augmentation has been shown to significantly improve image classification and object detection (Barret Zoph). Its potential has not been thoroughly investigated for human pose estimation. Radosavovic (18) investigated omni-supervised learning, a special regime of semi-supervised learning in which the learner exploits all available labeled data plus internet-scale sources of unlabeled data. In the omni-supervised learning, data distillation was firstly brought out as a simple strategy for human pose estimation, shown in
Figure 2.13. The idea is to generate annotations on unlabeled data using a model trained on large amounts of labeled data, and then retrain the model using the extra generated annotations.

Figure 2.13: Data Distillation (18). In data distillation, ensembled predictions from a single model applied to multiple transformations of an unlabeled image are used as automatically annotated data for training a student model.

2.3.9 Attention Mechanism

Attention mechanism is one component of a network's architecture and is in charge of managing and quantifying the interdependence. Attention mechanism in deep ConvNet is achieving big success in recent years. The attention mechanism used in human pose estimation is firstly brought out in Chu's work (51). Figure 2.14 shows how the multi-context mechanism improved the human pose estimation performance.

2.3.10 The Development Of Backbone ConvNet

Recently there are a lot of novel backbone ConvNet architecture design to aggregate feature maps in different scales for computer vision tasks such as object detection, image classification, image segmentation and human pose recognition. Ke et al. (Ke Sun) conducted a further study on high resolution representations by introducing a simple yet effective
Figure 2.14: Multi context attention mechanism (51). (a) is the original heatmap and predicted pose. (b) is the attention map. (c) is the refined pose by attention, the part attention model could fix the double counting problem.
modification and applied it to a wide range of vision tasks. Gao et al. proposed a novel building block for CNNs, namely Res2Net (Shang-Hua Gao), by constructing hierarchical residual-like connections within one single residual block. Yupeng Chen et al. proposed to factorize the mixed feature maps by their frequencies, and designed a novel Octave Convolution (Yunpeng Chen) operation to store and process feature maps that vary spatially slower at a lower spatial resolution reducing both memory and computation cost. Unlike existing multi-scale method-ods, Octave Convolution (Yunpeng Chen) is formulated as a single, generic, plug-and-play convolutional unit that can be used as a direct replacement of convolutions without any adjustments in the network architecture. The proposed deep layer aggregation structures by Fisher Yu (11) iteratively and hierarchically merged the feature hierarchy to make networks with better accuracy and fewer parameters. Their experiments across architectures and tasks showed that deep layer aggregation improves recognition and resolution compared to existing branching and merging schemes. Thus it is a promising direction to go for designing multigrid neural architecture of human pose recognition.

- **Octave Convolution**: Octave Convolution (Yunpeng Chen) proposes to factorize the mixed feature maps by their frequencies, and designs a novel octave convolution operation to store and process feature maps that vary spatially "slower" at a lower spatial resolution reducing both memory and computation cost.

- **Selective Kernel Network**: Selective kernel network (49) unit is designed, in which multiple branches with different kernel sizes are fused using softmax attention that is guided by the information in these branches.

## 2.4 The Proposed Work and Contributions

In summary, classical approach for human pose estimation has good representation of the pose structure such as pictorial structure (30) or flexible mixtures-of-parts (55). However, it lacks the good representation of features. Deep ConvNet has good representation of feature but no structure information. The comparision is shown in *table 2.1*. In order to achieve best performance, the proposed method is to have the best of both traditional method and Deep ConvNet, good features and good representation of pose.
Table 2.1: Comparision between traditional method and DeepConvNet

<table>
<thead>
<tr>
<th></th>
<th>Traditional method</th>
<th>DeepConvNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>Weak (handcrafted)</td>
<td>Good (end to end)</td>
</tr>
<tr>
<td>Representation of Pose</td>
<td>Good (structure)</td>
<td>No</td>
</tr>
</tbody>
</table>

In order to achieve this goal, two methods are explored to add head ConvNet on top of existing feature backbone. The first method is multi-branch heads with limb guided attraction field. The other method is light-head multi-kernel with self-attention mechanism. They are studied separately in this thesis in top-down pipeline.
A lot of research in 2D human pose estimation try to handle hard keypoints such as ankle and knee by introducing refined Convnet architecture (56). However, the thesis tries to solve this problem in a simpler way. The thesis makes the assumption that different body parts have different receptive field sizes. In order to achieve higher interpretable pose recovery results while remaining the good performance, this chapter proposes multi-kernel head with self-attention mechanism. The MS COCO dataset and evaluation metric. MS COCO dataset is widely used to evaluate the performance of 2D human pose estimation algorithm. There are open competitions every year on the MS COCO dataset. The dataset is more challenging than MPII dataset and it is currently used to evaluate the performance of human pose estimation algorithm.
3.1 Motivation

The current single-branch human pose estimation algorithm using a single-kernel head for learning keypoint heatmaps, shown in figure 3.1. The method may undergo strong assumptions that all keypoints use the same receptive field, and thus prevent the model from learning more accurate heatmaps. Intuitively we know different skeletal keypoints have different intrinsic scales and different surrounding contexts. Even for the same keypoint (e.g., ankle), we need to account for structural variations (e.g. frontal view vs profile view, self-occlusion) with different receptive fields. From the motivation, the proposed method adds a filter bank to generate heatmap bank. The filter bank has multiple kernels. The final heatmap is generated by the fusion of the heatmap bank. The overview of the proposed method is shown in figure 3.2.

![Figure 3.1: Single Kernel Head. The image firstly goest through ConvNet Feature Backbone such as SimpleBaseline (5) or HRNet (27), then map to the feature map with the number of the channels is the number of keypoints. The kernel size of the mapping is 3*3.](image)

3.2 Method Overview

This section discusses the proposed multi-kernel heads with self-attention and related study. Multi-kernel heads is discussed in the first section and fusion mechanism is discussed in the second section.
Figure 3.2: Motivation of Multi-Kernel Heads Design. Compared to Single Kernel head, the multi-kernel heads add a filter bank with multiple kernel sizes to generate heatmap bank. Then the final heatmap is generated by the fusion of the heatmap bank.

3.2.1 Multi-kernel Heads

The proposed multi-kernel heads adds multiple kernels on top of the ConvNet feature backbone such as HRNet (27). The heatmap bank has the regressed heatmaps for keypoints with the supporting kernel \((k_i, k_j)\). Figure 3.3 is the deep ConvNet architecture with multi-kernel head for human pose estimation. The sizes of kernel bank are set to be \((3, 3)\), \((3, 5)\), \((3, 7)\), \((5, 3)\), \((5, 5)\), \((5, 7)\), \((7, 3)\), \((7, 5)\) to \((7, 7)\) in the experiments.

Figure 3.3: Method Overview of Multi-Kernel Heads. The filter bank of the multi-kernel heads support 9 kernels. The widths are 3, 5 and 7 and the heights are 3, 5 and 7.

The plots of different keypoints combinations for the COCO 2017 train dataset are shown
in figure 3.4.

Figure 3.4: The plots of different human body keypoints combination. The middle point is in the origin. The blue line represents the left keypoint, the red line represents the right keypoint. Six combinations are listed in the figure. There are six combinations in the figure.
Take the first scatter plot from figure 3.4 for example, the middle keypoint (left eye) is in the origin, the red line represents the nose while the blue line represents left ear. The visualization of scatter plot validates the phenomena that human keypoint pairs far from the center have larger kernel size, while human keypoint pairs near the center have smaller kernel size. From this observation, we could make the assumption that different body parts may select different kernel sizes.

### 3.2.2 Fusion

The fusion mechanism is added to the heatmap bank, because the heatmap in the heatmap bank is not directly comparable due to using different kernels, as shown in figure 3.5. Then the self-attention mechanism is added to the heatmap bank to recalculate the relative importance of each heat map in the heatmap bank. In the last step, the final heatmap is the mean of recalibrated heatmaps.

![Figure 3.5: Fusion of Multi-kernel Heads](image)

Figure 3.5: Fusion of Multi-kernel Heads. Firstly self-attention mechanism works in the heatmap bank to recalibrate the relative importance of each heatmap, then the mean operation is adapted to get the final heatmap.
3.2.3 Self-Attention

Visual attention is an essential mechanism of the human brain to understand scenes effectively. Attention mechanism is widely used in computer vision tasks such as classification, segmentation and object detection. Chu (51) proposes to incorporate convolutional neural network with the multi-context attention mechanism into an end-to-end framework for human pose estimation. In their work, three types of attentions are designed, multi-resolution attention within each hourglass, multi-semantics attention across several stacks of hourglass, and a hierarchical visual attention scheme to zoom in on local regions to see clearer. This is the first work to investigate the attention mechanism with deep ConvNet architecture for 2D human pose estimation.

Hu etc. proposes the "Squeeze-and-Excitation" (SE) block (21) which could adaptively recalibrate channel-wise feature responses by explicitly modelling interdependencies between channels. The squeeze and excitation blocks bring significant improvements in performance of ILSVRC 2017 classification task for existing state-of-the-art deep ConvNet at slight additional computational cost. It achieves about 25 percent improvement surpassing the winning entry of 2016 (6).

The self-attention mechanism adds a light-weight branch into existing architecture, it has global pooling, two fully connected layers to get the importance weight of each channel in the heatmap bank. The reduction ratio $r$ introduced in the two fully connected layers is a hyperparameter to vary the capacity and computational cost of the SE blocks. It is set to be 1 in the thesis because Hu (21) concludes that the performance is not affected much by a range of reduction ratios. Then the self-attention mechanism recalibrates the relative importance of each heatmap in the heatmap bank, shown in figure 3.6.

3.3 Experiments

The experiments are conducted on MS COCO Dataset. HRNet is used as the ConvNet feature backbone (27). The multi-kernel heads with self-attention mechanism is added on top of the backbone HRNet (27). The self-attention mechanism is added to the multi-kernel heads to recalibrate the heatmaps in heatmap bank. In the last, the mean operation is conducted to fuse all the heatmaps in the heatmap bank.

The learning schedule follows the same setting. The base learning rate is set as $1e^{-3}$, and is dropped to $1e^{-4}$ and $1e^{-5}$ at the 170th and 200th epochs. The training process
Figure 3.6: Self-attention Mechanism. The importance weight is calculated by adding global pooling and two fully connected layers to the heatmap bank. Then each heatmap in the heatmap bank is multiplied the weight.

Table 3.1: Performance analysis on MS COCO dataset on AP56 bounding box

<table>
<thead>
<tr>
<th>Percentage of train data</th>
<th>AP</th>
<th>AP_{50}</th>
<th>AP_{75}</th>
<th>AP_{M}</th>
<th>AP_{L}</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRNet</td>
<td>75.8</td>
<td>90.6</td>
<td>82.5</td>
<td>72.0</td>
<td>82.7</td>
<td>80.9</td>
</tr>
<tr>
<td>proposed method</td>
<td>76.4</td>
<td>90.6</td>
<td>82.9</td>
<td>72.6</td>
<td>83.3</td>
<td>82.0</td>
</tr>
</tbody>
</table>

is terminated at 240 epochs. In the inference stage, the heatmap is averaged over the headmaps of the original and flipped images.

3.3.1 Performance on COCO Val2017 Dataset

The performance on COCO val2017 dataset with detector having human AP of 56.4 on COCO val2017 dataset is shown in table 3.1. The performance on COCO val2017 dataset with ground-truth bounding boxes is shown in table 3.2.

The pose estimation result of several images from COCO val2017 dataset is shown in figure 3.7.

Table 3.2: Performance analysis on MS COCO dataset on groundtruth bounding box

<table>
<thead>
<tr>
<th>Percentage of train data</th>
<th>AP</th>
<th>AP_{50}</th>
<th>AP_{75}</th>
<th>AP_{M}</th>
<th>AP_{L}</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRNet</td>
<td>77.7</td>
<td>93.6</td>
<td>84.7</td>
<td>74.8</td>
<td>82.5</td>
<td>80.4</td>
</tr>
<tr>
<td>proposed method</td>
<td>78.5</td>
<td>93.6</td>
<td>85.7</td>
<td>75.6</td>
<td>83.2</td>
<td>81.1</td>
</tr>
</tbody>
</table>
3.3.2 Visualization of multi-scale feature maps

In this section, we will discuss the visualization of multi-kernel feature maps. figure 3.8 shows a man with left ankle occluded by a dog. The figure is from COCO 2017val dataset which has 5000 images.

The feature maps in feature bank generated by different kernel size is shown in figure 3.9. It validates the assumption made in the motivation that using a single-kernel head for learning keypoint heatmaps may undergo strong assumptions, and thus prevent the model from learning more accurate heatmaps.

3.3.3 Error Diagnosis on COCO Val2017 Dataset

To take a closer look at the errors on MS COCO dataset, the paper utilizes the coco-analyze tool developed by Ronchi. Rochi (31) et al. propose a new method to analyze the impact of errors in algorithms for multi-instance pose estimation. The technique is applied to compare the two leading methods(OpenPose (59) and GRMI (55)) for human pose estimation.
algorithm on MS COCO dataset. There are several types of errors discussed in the following section. Localization is the poor predicted location of the keypoints belonging to a detected instance. Scoring is sub-optimal confidence score assignment. A scoring error occurs when two detections are near a ground-truth annotation and the one with the highest confidence has the lowest OKS. It happens when there are overlapping people and detections. Background False Positives are detections without a ground-truth annotation match. False Negatives are missed detections. The fine-grained precision-recall curves are obtained by fixing an OKS threshold and evaluating the performance of an algorithm after progressively correcting its mistakes (32). The ground-truth instances in the MS COCO dataset is separated into twelve benchmarks, based on number of visible keypoints and overlap between annotations. The performance and error sensitivity to occlusion and crowding is shown in figure 3.10 and figure 3.11. The proposed multi-kernel heads with self-attention mechanism method could enhance the performance compared to HRNet (27).

- **Original Dts**: PR obtained with the original detections at OKS=.75 (AP at strict OKS), area under curve corresponds to APOKS=.75 metric.;

- **Localization error**: Miss, Swap, Inversion, Jitter.

- **Opt. Score**: PR at OKS=.75(AP at strict OKS), after all the algorithm's detections have been rescoring using an oracle function computed at evaluation time. As a result of
Figure 3.9: Visualization of Feature Maps in Feature Bank. The feature map generated by 9 different kernel sizes is shown in the figure. The size of the rectangle represents the size of the kernel to generate feature map.
Figure 3.10: Performance and Error Sensitivity to Occlusion and Crowding for HRNet (27). The PR curves showing the performance of obtained by progressively correcting errors of each type at the OKS evaluation threshold of .75 on the twelve Occlusion and Crowding Benchmarks.

- The rescoring the number of matches between detections and ground-truth instances is maximized.

- **FP:** PR after all background fps are removed. FP is a step function that is 1 until max recall is reached then drops to 0 (the curve is smoother after averaging across categories).

- **FN:** PR after all remaining errors are removed (trivially AP=1).

Among these four types of errors (the localization error, scoring error, background false positives and false negatives), the research focus of the proposed method is localization error. A localization error often occurs when the location of the keypoints in a detection
Figure 3.11: Performance and Error Sensitivity to Occlusion and Crowding for Multi-kernel heads with self-attention mechanism. the PR curves showing the performance of obtained by progressively correcting errors of each type at the OKS evaluation threshold of .75 on the twelve Occlusion and Crowding Benchmarks.
results in an OKS score with the corresponding ground-truth match that is lower than the evaluation threshold. They are typically due to the fact that body parts are difficult to detect because of self occlusion or occlusion by other objects. The different types of localization error are discussed in the following section and the visualization of some samples are shown in figure 3.12.

- **Jitter**: small error around the correct keypoint location, that is the object similarity between predicted keypoints and groundtruth keypoints are between 0.5 and 0.85.

- **Miss**: large localization error, the detected keypoint is not within the proximity of any body part, that is the object similarity between predicted keypoints and groundtruth keypoints is less than 0.5.

- **Inversion**: confusion between semantically similar parts belonging to the same instance. The detection is in the proximity of the true keypoint location of the wrong body part.

- **Swap**: confusion between semantically similar parts of different instances. The detection is within the proximity of a body part belonging to a different person.

The comparison of localization error between HRNet (27) and the proposed method of multi-kernel head with attention mechanism is shown in Figure 3.13. It shows that the proposed multi-kernel head with attention mechanism could reduce the three categories of error: jitter, miss and inversion.
Good means that the object similarity between predicted keypoints and groundtruth keypoints is less than 0.5. The proposed multi-head heads with self-attention mechanism get better performance.

The inversion localization error distribution in different human body parts between HRNet (27) and the proposed multi-kernel heads with self-attention mechanism is shown in Figure 3.14. The multi-kernel heads with self-attention mechanism method could reduce inversion localization error in wrists, eyes, knees, ankles.
The jitter localization error distribution of the four types of error in different human body parts between HRNet (27) and the proposed multi-kernel heads with self-attention mechanism method are shown in figure 3.15. The multi-kernel heads with self-attention mechanism method could reduce jitter localization error in nose, eyes, ears, shoulder and elbows.
The miss localization error distribution of the four types of error in different human body parts between HRNet (27) and the proposed multi-kernel heads with self-attention mechanism method is shown in figure 3.16. The multi-kernel heads with self-attention mechanism method could reduce miss localization error in nose, wrists, eyes, ears and shoulders.
The swap localization error distribution of the four types of error in different human body parts between HRNet (27) and the proposed multi-kernel with attention mechanism method is shown in Figure 3.17. The multi-kernel heads with self-attention mechanism method could reduce swap localization error in knees, shoulder, ankles and elbows.
3.4 Summary

In summary, this chapter proposes multi-kernel heads with self-attention mechanism in top-down pipeline for 2D human pose estimation algorithm. The experiments of the second method are conducted on COCO dataset. It could achieve better performance on top of state-of-the-art backbone ConvNet featuremap such as HRNet (27). It could give more interpretable intermediate visualization result in some cases such as occlusion.
Chapter 1 gives the problem definition of 2D human pose estimation and its challenges. It also gives the overview of four benchmarks: Frames Labeled In Cinema (FLIC) Benchmark, Leeds Sports Pose Benchmark, MPII Benchmark and Common Objects in Context Dataset (COCO) Benchmark. Chapter 2 gives a detailed literature review for deep ConvNet architecture design for single-person human pose estimation, and it summarizes the contribution of the thesis. The proposed methods in the thesis are based on general top-down strategy. And this chapter focuses on the exploration to integrate structural information such as part affinity map and affinity graph map of limb with state-of-the-art convNet feature backbone such as HRNet (27) for single person pose estimation.
4.1 Motivation

Structural Information (such as Part Affinity Field, PAF) is exploited in the bottom-up pipelines for multi-person pose estimation, but is usually ignored in top-down pipelines although applicable. This chapter proposes the multi-stage multi-branch head that encodes the part affinity field and attraction field map information. The motivation of multi-stage multi-branch heads in top-down pipeline for multi-person pose estimation is shown in figure 3.1.

![Figure 4.1: Motivation for Multi-stage Multi-branch Heads. In the top-down pipeline, the proposed heads encode structural information such as local structural information and global structural information besides heatmap branch.](image)

4.2 Structural Information for Limb

Before deep ConvNet Era, pictorial model (30) and related work (55) dominate the 2D human pose estimation problem. In the pictorial model (30), \( m_i(l_i) \) is a function measuring the degree of mismatch when part \( v_i \) is placed at location \( l_i \) in the image. For a given pair of connected parts, \( d_{ij}(l_i, l_j) \) is a function measuring the degree of deformation of the model when part \( v_i \) is placed at location \( l_i \) and part \( v_j \) is placed at location \( l_j \). Then the optimal
match of the model to the image is naturally defined as

\[ L^* = \arg \min_L \left( \sum_{i=1}^{n} m_i(l_i) + \sum_{v_i,v_j \in E} d_{ij}(l_i,l_j) \right) \]  (4.1)

It is the configuration minimizing the sum of the match costs \( m_i \) for each part and the deformation costs \( d_{ij} \) for connected pairs of parts (37). In those work, spatial relation is modeled by the edges in tree structure. Each connection \((v_i, v_j)\) is characterized by the ideal relative location of the two connected parts \( s_{ij} \), and a full covariance matrix \( \Sigma_{ij} \) which in some sense corresponds to the stiffness of the spring connecting the two parts. So the connection parameters are \( c_{ij} = (s_{ij}, \Sigma_{ij}) \). We model the distribution of the relative location of part \( v_i \) with respect to the location of part \( v_j \) as a Gaussian distribution with mean \( s_{ij} \) and covariance \( \Sigma_{ij} \). The maximum likelihood parameters of these spatial distributions for pairs of parts can be estimated using training examples.

Another obvious constraint is given by a collection of joint coordinates (38), \( X = [p_1, \ldots, p_n] \), where \( p_j = [x_j, y_j, z_j]^T \) are the coordinates of the \( j \)-th joint. Presumably the advantage is that it allows one to work on the euclidean space as opposed to working on the joint angle space. However, the kinematic constraints such as bone-length preservation and connected parts have to be imposed during optimization. In the proposed method, the local structural information is represented by a deep ConvNet. There are three categories of structural information brought out in the proposed method: part affinity map and attraction field map. They are utilized inherently to enhance the performance of human pose estimation algorithm.

### 4.2.1 Part Affinity Field for Limb

Part affinity field (PAF) is first proposed by Zhe (59) to learn to associate body parts with individuals in multi-person human pose recognition. The original architecture encodes global context, allowing a greedy bottom-up parsing step that maintains high accuracy while achieving realtime performance, irrespective of the number of people in the image. The part affinity field is a 2D vector field. For each pixel in the area belonging to a particular limb, a 2D vector encodes the direction that points from one endpoint of the limb to the other. Each type of limb has a corresponding affinity field joining its two associated body parts. In the proposed method, it is supposed to introduce more context information for
limb and improve single human pose estimation algorithm performance.

Consider a single limb, let $x_{j_1,k}$ and $x_{j_2,k}$ be the ground truth positions of body parts $j_1$ and $j_2$ from the limb $c$ for person $k$ in the image. If a point $p$ lies on the limb, the value at $L^*_{(c,k)}(p)$ is a unit vector which points from $j_1$ to $j_2$, and is equal to $(x_{j_2,k} - x_{j_1,k}) / \| x_{j_2,k} - x_{j_1,k} \|_2$; for the rest of points, the vector is zero-valued.

### 4.2.2 Attraction Field Map for Limb

Xue et al. proposed a region-partition based attraction field (35) dual representation for line segment maps, and thus posed the problem of line segment detection (LSD) (40) as a region coloring problem. Line segment detection usually consists of two steps: line heat map generation and line segment model fitting. The proposed attraction field map by Xue consists of three components.

- **Region-Partition Map**: A region-partition map is the map in which every pixel is assigned to one and only one line segment.

- **Attraction Field Map**: An attraction field map is a map in which every pixel in a partition region is encoded by its 2D projection vector w.r.t. the associated line segment.

- **Squeeze Module**: A squeeze module is a module which squashes the attraction field to a line segment map that almost perfectly recovers the input one.

Every pixel in the image will be assigned to one line segment. Consider a pixel $p$ and a line segment $l_i = (x_{i,s}^e, x_{i,e}^e)$, the pixel $p$ is first projected to the straight line going through $l_i$ in the continuous geometry space. If the projection point is not on the line segment, the closest end-point of the line segment is used as the projection point. The pixel assignment method is shown in [figure 4.2](#).

In the proposed single pose estimation algorithm, the attraction field map is utilized in the limb region. A light-weight squeeze module is not needed in the deep ConvNet design. In the proposed method, the fitted region is considered to be the line segment around the line. The 2D attraction or projection vector for a pixel $p$ can be defined as [equation 3.2](#).

$$a(p) = p^* - p$$  \hspace{1cm} (4.2)

The projected vector is a two-dimensional vector, and the fitted $d_x$ map and $d_y$ map are shown in [Figure 4.3](#).
Figure 4.2: Pixel Assignment Method. Every pixel in the entire image is attracted to one and only one limb based on the distance.

Figure 4.3: Attraction Field Map (35). The left image is the attraction field map for $d_x$, the right image is the attraction field map for $d_y$.

The line segments are constructed by connecting two body part keypoints of all the keypoints. In the experimental test cases, all line segment maps are first converted in the training dataset to their attraction field maps. Then, the ConvNet is trained end-to-end to predict the attraction field maps from raw input images. HRNet (27) is used as the base backbone network. After the attraction field map is computed, the squeeze module is used
to compute its line segment map. The result of Attraction Field Representation (35) of FLIC dataset based on HRNet (27) was shown in figure 3.4.

Figure 4.4: Line segment detection for 2D Human Pose. The left image is from FLIC test dataset. The right image is the fitted line segment in testing image.

4.3 Method Overview

The proposed method is to adapt top-down approach for the multi-person pose estimation. And the key problem of the proposed method is how to design a better deep ConvNet architecture combining structural information for single-person pose estimation. Researcher study a lot about how to fuse feature maps at different scales (27) (1) (56), while little research is conducted on how to combine the structural information for single-person pose estimation. There is some research trying to combine probability graphical model with deep ConvNets (23), while the proposed method uses the deep ConvNets to directly represent the structure information.

The proposed single person pose estimation architecture used HRNet (27) as the base network and utilized three stages of refinement: vanilla heatmap, part affinity map and attraction field map for limb. Also, these three lower-level feature maps were added to the
higher feature maps in order to get better representation.

The input of the proposed method is a color image of size $W \times H$ and the output is the 2D locations of anatomical keypoints for each person. A feedforward network simultaneously predicts a set of 2D confidence heatmaps $f_H$ of body part locations, a set of 2D vector fields $f_P$ of part affinity map, a set of 2D vector fields $f_A$ of attraction field map. The proposed multi-branch method utilized deep ConvNet architecture for human pose recognition was shown in figure 4.5. The proposed architecture simultaneously predicts confidence heat maps, part affinity fields and attraction field maps which encode part-to-part association in intermediate stages, while the last stage only outputs confidence heatmaps.

Let $W$ be a binary mask with $W(p) = 0$ when the annotation is not used at an image location $p$, otherwise $W(p) = 1$. $i$ is the channel number for specific type of keypoints. $H(i)$ is the groundtruth heatmap for ith keypoint, and $H^*(i)$ is the predicted heatmap. $P(i)$ is the groundtruth part affinity graph for ith keypoint, and $P^*(i)$ is the predicted part affinity graph. $A(i)$ is the groundtruth attraction field map for ith keypoint, and $A^*(i)$ is the predicted attraction field map. The total loss is the sum of euclidean distance loss of vanilla heatmap, part affinity graph and attraction field map can be written as equations (4.5) to (4.8)

\[
\begin{align*}
  f_H &= \sum_i W(i) \cdot ||H(i) - H^*(i)||^2 \quad (4.3) \\
  f_P &= \sum_i W(i) \cdot ||P(i) - P^*(i)||^2 \quad (4.4)
\end{align*}
\]
\[ f_A = \sum_i W(i) \cdot \text{smoothl1loss}(A(i), A^*(i)) \] (4.5)

The intermediate is applied at each stage to address the vanishing gradient problem by replenishing the gradient periodically. The overall objective is as equation 3.6, \( c \) is the factor to control the ratio between attraction field map and heatmap/part affinity map.

\[ f = f_H + f_P + c f_A \] (4.6)

### 4.4 Experiments

In the thesis, the experiments of multi-branch heads with limb guided attraction field are conducted in FLIC dataset and MPII dataset. The experiments are described in the following sections.

#### 4.4.1 Experiments on FLIC Dataset

The training data is augmented by random rotation ([45°, 45°]) and Adam optimizer is adapted. The learning rate schedule follows the same setting for all the experiments: the base learning rate is set as \( 1e^{-3} \). The training process is terminated at 100 epochs, it is less than that of HRNet (27) because FLIC is a smaller dataset compared with MPII dataset and MS COCO dataset. The pose estimation result of several images from FLIC test dataset is shown in figure 4.6.
Table 4.1: FLIC results (PCK@0.2)

<table>
<thead>
<tr>
<th>Method</th>
<th>Elbow</th>
<th>Wrist</th>
</tr>
</thead>
<tbody>
<tr>
<td>StackedHourglass</td>
<td>99.0</td>
<td>97.0</td>
</tr>
<tr>
<td>3-level DLCM</td>
<td>99.5</td>
<td>98.5</td>
</tr>
<tr>
<td>HRNet-W32 Reproduced</td>
<td>98.6</td>
<td>96.7</td>
</tr>
<tr>
<td>Proposed method</td>
<td><strong>99.9</strong></td>
<td><strong>99.8</strong></td>
</tr>
</tbody>
</table>

Figure 4.6: Pose Estimation Result on FLIC test dataset. The green circle represents ground truth keypoints while the red plot represents the predicted keypoints.

The performance of PCK@0.2 on FLIC test dataset is reported in table 3.1. The proposed multi-branch heads with limb guided attraction field could achieve state-of-the-art performance of single-person human pose estimation in FLIC dataset comparable with StackedHourglass (1) and CompositionalModel (11).

4.4.2 Experiments on MPII Dataset

The data augmentation includes random rotation ([45°, 45°], the Adam optimizer is adapted. The learning schedule follows the same setting. The base learning rate is set as 1e−3, and is dropped to 1e−4 and 1e−5 at the 170th and 200th epochs, respectively. The training
Table 4.2: Performance of PCKh@0.1 on MPII validation dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Shoulder</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRNet-W32</td>
<td>51.1</td>
<td>42.7</td>
<td>42.0</td>
<td>41.6</td>
<td>17.9</td>
<td>29.9</td>
<td>31.0</td>
<td>37.7</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>50.5</td>
<td>43.9</td>
<td>44.6</td>
<td>43.0</td>
<td>18.9</td>
<td>30.5</td>
<td>33.1</td>
<td>39.1</td>
</tr>
</tbody>
</table>

process is terminated at 210 epochs, following the same design of HRNet (27). The heatmap is averaged over the headmaps of the original and flipped images. The pose estimation result on sample images of COCO val2017 dataset is shown in figure 4.7.

Figure 4.7: Pose Estimation Result on MPII validation dataset. The blue rectangles represent groundtruth while the red represents the predicted keypoint.

The performance of PCKh@0.1 on MPII validation dataset is shown in table 3.2, and the performance of PCKh@0.5 on MPII validation dataset is shown in table 4.3. The performance is comparable to state-of-the-art algorithm both in PCKh@0.1 and PCKh@0.5. The study demonstrates that structural information is beneficial in enhancing the performance of base ConvNet feature map.

In order to evaluate the performance of the proposed multi-kernel heads with self-
Table 4.3: Performance of PCKh@0.5 on MPII validation dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Shoulder</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRNet-W32</td>
<td>97.2</td>
<td>95.8</td>
<td>90.8</td>
<td>86.2</td>
<td>89.2</td>
<td>86.1</td>
<td>82.8</td>
<td>90.2</td>
</tr>
<tr>
<td>Proposed method</td>
<td>97.3</td>
<td>96.1</td>
<td>91.0</td>
<td>86.3</td>
<td>89.1</td>
<td>86.7</td>
<td>83.3</td>
<td>90.4</td>
</tr>
</tbody>
</table>

attention mechanism in different human body parts, the performance plot is shown in figure 4.8. The proposed multi-branch heads could achieve good performance in different body parts.

Figure 4.8: Performance Plot on MPII validation dataset for the proposed method. The are seven parts listed in the figure. The performance is worse in hard keypoints such as hip and ankle.
To evaluate the performance with or without structural information, the comparison between HRNet (27) and the proposed multi-branch heads with limb guided attraction field vector method on some validation images is shown in figure 4.9. The input size of image is 256 * 256. The left side is the result of HRNet (27) and the right side is the result of the proposed multi-branch heads method. The multi-branch heads method could fix some assignment error of the validation images.

Figure 4.9: Comparison between HRNet and the proposed heads. (a) gets the wrong location of the head with similar appearance. (b) fixes the error. (c) is double counting left ankle. (d) fixes the error.
There are some errors that the proposed multi-heads method could not fix such as occlusion, including self-occlusion or occlusion by other stuff, as shown in Figure 4.10.

![Figure 4.10: Mistakes of the proposed multi-branch heads method.](image)

- (a) ground truth pose
- (b) Proposed multi-branch heads
- (c) ground truth pose
- (d) Proposed multi-branch heads

There are also missing errors that the proposed multi-heads method could not fix, as shown in Figure 4.11.
4.5 Summary

This chapter proposes multi-branch heads with limb guided attraction field on top of state-of-the-art backbone ConvNet feature. The multi-stage idea is also adapted in single person pose estimation architecture such as Convolutional Pose Machines (56) and Stacked Hourglass (1). Multi-branch idea is exploited in bottom-up multi-person pose estimation algorithm. The proposed method combined multi-stage multi-branch design with structural information. The experiments are conducted in FLIC and MPII datasets. It could improve the performance with extra parameters.
5

SUMMARY AND FUTURE WORK

5.1 Summary

In the first chapter, the thesis gives the problem definition and introduces challenges with application of 2D human pose estimation algorithm. The thesis does the literature review in the second chapter. It discusses top-down and bottom-up approach for multi-person pose estimation. Top-down approach first detects the human bounding box and then locate the human body keypoints. Bottom-up approach firstly detects the human body parts and then assemble the parts. The thesis also introduces classical approach and deep ConvNet approach for single-person human pose estimation. It summarizes the shortcoming and strength of both methods. Classical approach lacks the good representation of features. Deep ConvNet approach has good representation of feature but no information of pose structure. In the literature review of the second chapter, the thesis discusses the research progress in deep ConvNet architecture design for single-person human pose estimation. The DeepPose is the first deep ConvNet architecture to solve 2D human pose estimation problem. Convolutional pose machines (42) is the first method to use heatmap as the rep-
presentation of keypoints. Stacked Hourglass (1) shows how repeated bottom-up, top-down processing used in conjunction with intermediate supervision is critical to improving the performance for 2D human pose estimation. Cascaded Pyramid Network (56) and the following work (15) discuss how to improve the performance of 2D human pose estimation by the multi-stage deep ConvNet architecture. Simple Baseline (5) comes up with a simple way to obtain high resolution feature map (thus keypoint heatmaps) by three upsampling steps and three levels of non-linearity. Deep high resolution feature map (27) is able to maintain the high resolution instead of recovering the resolution through a low-to-high process as it has high-to-low resolution subnetworks in parallel rather than in series as done in most existing solutions. The literature review part also gives brief introduction of other related topics in human pose estimation: pose refinement, self-supervised learning, attention mechanism and the development of backbone convNet such as octave convolution (Yunpeng Chen) and selective kernel network (49).

The main part is the chapter 3 and chapter 4 of the thesis. The chapter 3 and chapter 4 propose two methods based on deep ConvNet architecture to conduct human pose estimation in top-down pipeline. The first method is multi-branch heads with limb guided attraction field. The multi-branch includes local structural information (part affinity graph) and global structural information (attraction field map) on top of state-of-the-art backbone feature map (e.g. HRNet (27)). The experiments are conducted in MPII dataset and FLIC dataset. The first method could achieve state-of-the-art performance such as HRNet (27) with extra parameters introduced. The second method is to add multi-kernel heads with self-attention mechanism on top of state-of-the-art backbone feature map (e.g. HRNet (27)). The experiments of the second method are conducted on COCO dataset. The multi-kernel heads with self-attention mechanism could improve the performance with more interpretable pose recovery.

5.2 Future Work

There are several directions to go for improving the performance of pose estimation by integrating deep ConvNet and structural information.

- **Improvements in network design:** There are minor improvements in the network design such as how to integrate the low-resolution feature map in later refine stage.
• **Coarse-to-refine**: Integrate the coarse-to-refine like a refine network for hard keypoints like knee and ankle should improve the performance on MPII dataset.

• **How to integrate feature map with different scales**: There are other base neural networks worth integrating feature maps with different scales. For example, there is one promising direction to use deep aggregation network with OctConv (Yunpeng Chen).

• **Novel design of heatmap**: The fusion of a confidence map with a regression for keypoint detection was introduced in (13). The Part Intensity Fields (PIF) detect and precisely localize body parts. They are composed of a scalar component for confidence, a vector component that points to the closest body part of the particular type and another scalar component for the size of the joint. Feng Zhang etc. (10) finds that the process of decoding the predicted heatmaps into the final joint coordinates in the original image space is surprisingly significant for human pose estimation performance, which nevertheless is not recognised before., they further probes the design limitations of the standard coordinate decoding method widely used by existing methods, and proposes a more principled distribution-aware decoding method.

The demo of the parsing graph is shown in **figure 5.1**. All of the human keypoints represent a tree, and the limb could decide how to select different kernel size. As in the graph, the kernel size of the nose is 3*3, the kernel size of the left eye is 7*7, the kernel size of the right eye is 3*3.
Based on the multi-scale head with attention mechanism method proposed in Chapter 4, there is one promising design how to integrate the multi-scale information with part affinity map in human pose estimation algorithm. The proposed new architecture is shown in Figure 5.2. There are two branches in the head of HRNet (27): the multi-kernel keypoint head and the part affinity graph head. The kernel is selected by combining the heatmap score and part affinity graph response score. The refined loss could also be added after the dynamic programming process.

Figure 5.1: Parsing graph of human pose the keypoints kernel size is selected by combining information of heatmap and part affinity graph.
Figure 5.2: Proposed new architecture for human pose estimation. There are two branches: heatmap and part affinity graph branch. $l_{paf}$ is computed as in OpenPose (59) and multi-stage multi-branch heads. The inference is done by dynamic programming. There is also a refine stage to handle hard keypoints by $l_{dp}$ loss.
BIBLIOGRAPHY


67


PERFORMANCE ON FLIC DATASET
Figure 3: Performance Plots on FLIC dataset Combining Heatmap with Offset
Figure 4: Performance Plots on FLIC dataset Combining Heatmap and Attraction Field Map
Figure 5: Performance Plots on FLIC dataset Combining Heatmap and Part Affinity Map