简介

这篇文章的主要motivation在于多人姿态检测时，SSPE（single-person pose estimator）对bbox的准确性要求很严格，如果bbox预估不准确那么最后pose估计会非常不准。所以作者提出了RMPE(regional multi-person poes estimation)来解决这个问题。
主要贡献有三个：

1. Symmetric Spatial Transformer Network(SSTN)
2. Parametric Pose Non-Maximum-Suppersion(NMS)
3. Pose-Guided Proposaks Generator (PGPG)

之前pipline的主要问题是：

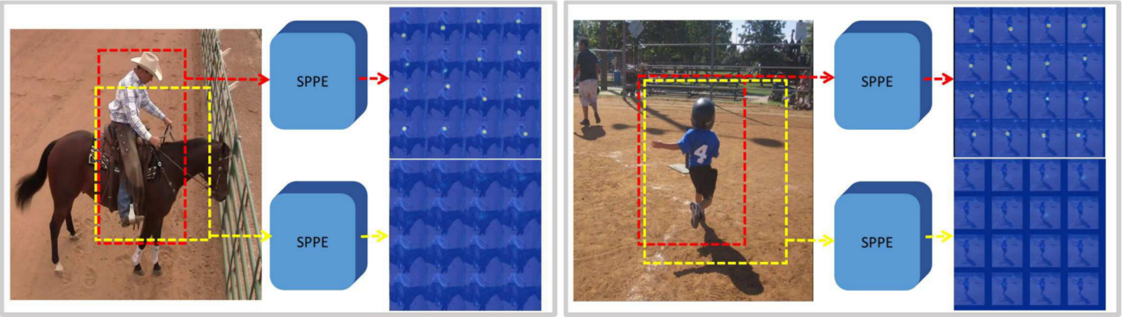


Figure 1. Problem of bounding box localization errors. The red boxes are the ground truth bounding boxes, and the yellow boxes are detected bounding boxes with $IoU > 0.5$. The heatmaps are the outputs of SPPE [28] corresponding to the two types of boxes. The corresponding body parts are not detected in the heatmaps of the yellow boxes. Note that with $IoU > 0.5$, the yellow boxes are considered as "correct" detections. However, human poses are not detected even with the "correct" bounding boxes.

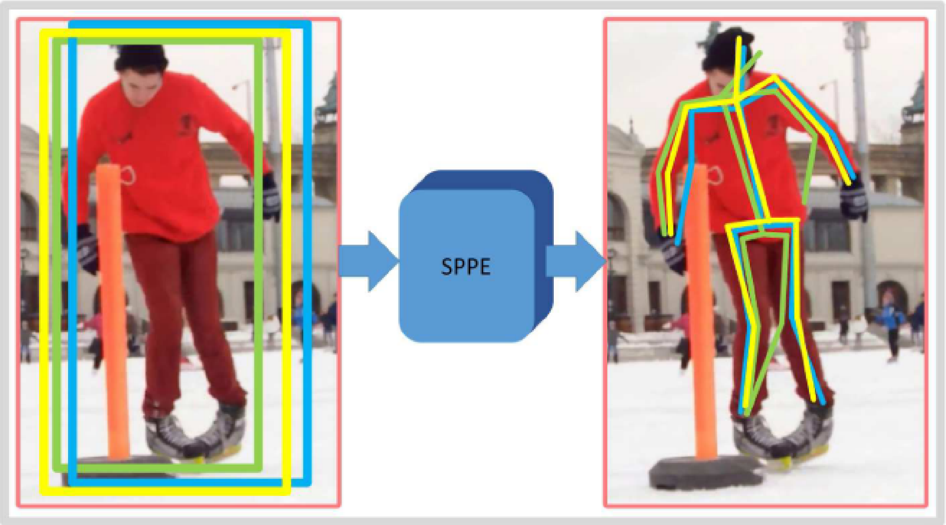


Figure 2. Problem of redundant human detections. The left image shows the detected bounding boxes; the right image shows the estimated human poses. Because each bounding box is operated on independently, multiple poses are detected for a single person.

RMPE

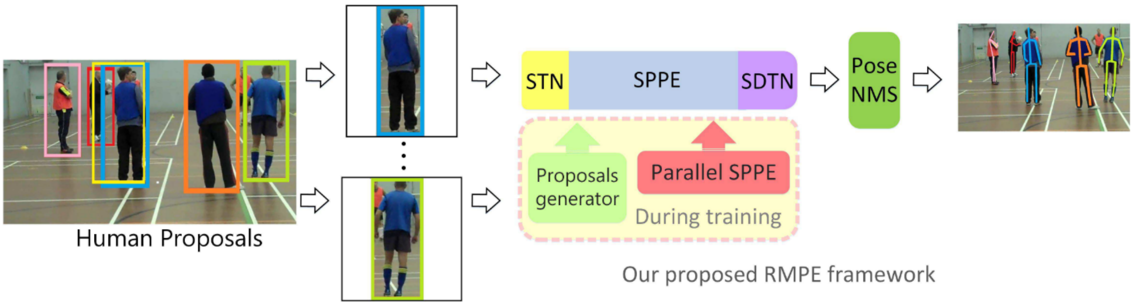


Figure 3. Pipeline of our RMPE framework. Our **Symmetric STN** consists of **STN** and **SDTN** which are attached before and after the SPPE. The **STN** receives human proposals and the **SDTN** generates pose proposals. The **Parallel SPPE** acts as an extra regularizer during the training phase. Finally, the **parametric Pose NMS (p-Pose NMS)** is carried out to eliminate redundant pose estimations. Unlike traditional training, we train the SSTN+SPPE module with images generated by **PGPG**.

SSTN 和 Parrallel SPPE

为了解决bbox的漂移问题，作者提出用SSTN + Parrallel SPPE解决。

作者是想用STN自适应的提取出human proposal。
再利用SDTN(detransformer)将估计的human pose再反映射回原来的坐标系当中。我的理解是STN是在SPPE的输入，也就是crop好的img上进行的，再通过SDTN把最后的结果反映射回去，得到最终的结果。

Parallel SPPE

这个模块的主要作用是帮助STN学习到更好的特征。加了一个平行的SPPE模块，两个模块的STN一致但是这个模块没有SDTN。这个模块的label是直接centered的，而不是detector的结果。
训练的时候这个parallel SPPE直接freeze，这个模块只是反向传播center-located pose erros给STN模块，帮助STN学习到centered feature。测试的时候这模块不要

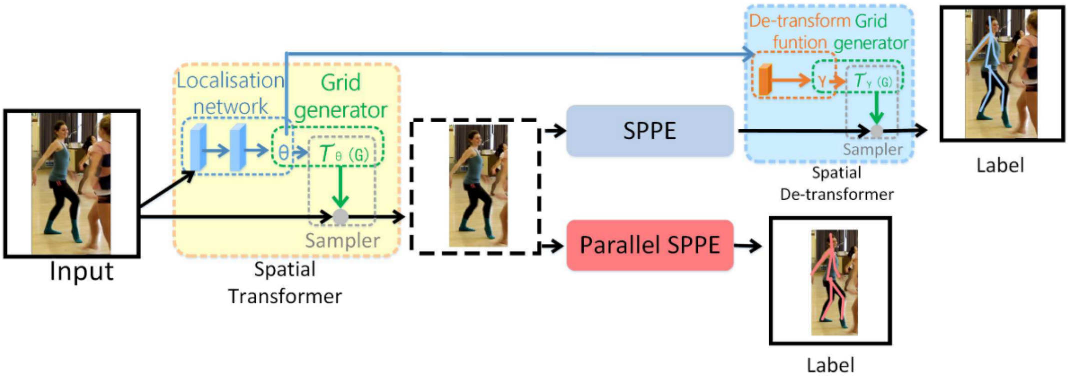


Figure 4. An illustration of our symmetric STN architecture and our training strategy with parallel SPPE. The STN used was developed by Jaderberg *et al.* [22]. Our SDTN takes a parameter $\boldsymbol{\theta}$, generated by the localization net and computes the γ for de-transformation. We follow the grid generator and sampler [22] to extract a human-dominant region. For our parallel SPPE branch, a center-located pose label is specified. We freeze the weights of all layers of the parallel SPPE to encourage the STN to extract a dominant single person proposal.

这里其实相当于一个regularizer。这个模块似乎可以通过一个center-located pose regression loss得到，在SPPE之后SDTN之前，但是这样会影响SPPE的性能。

Parametric Pose NMS

提出了基于pose的一套评分用于NMS，这里和`pose flow`里一致

PGPG（Pose-guded Proposals Generator)

基于two-stage不完美的检测框，作者想通过gt box去生成这些bbox，所以用一个学习分布规律的方法，去学习bbox的offset distrubution然后从这个分布里sample，因为不同姿势的分布是不一样的，所以作者先将姿势通过atomic pose进行划分。这个分布已`Gaussian mixture`作为先验假设

实验

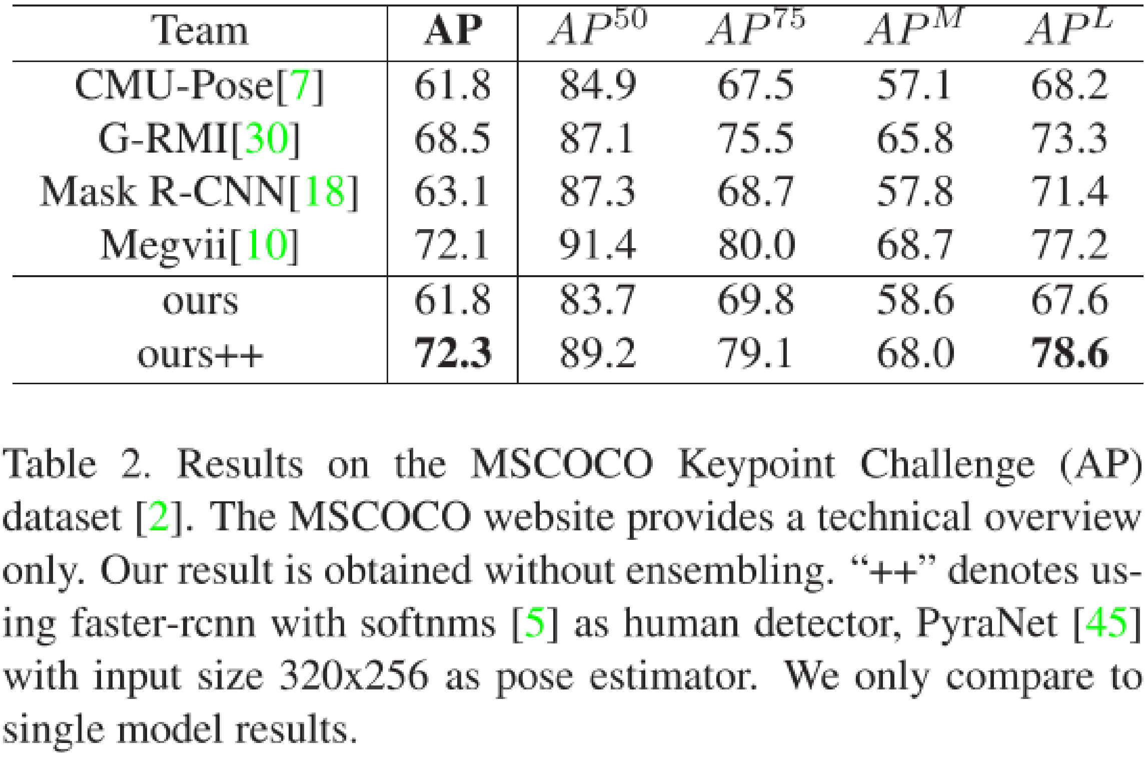
实验数据用的MPII和COCO,coco有超过10w human instances作为训练集 。

测试细节

- VGG-based SSD-512
- 检测到的bbox都沿着宽和高各扩30%的长度。
- 4-stack hourglass作为SPPE
- ResNet-18作为STN

还做了一个resnet152 based faster-RCNN + PyraNet的版本

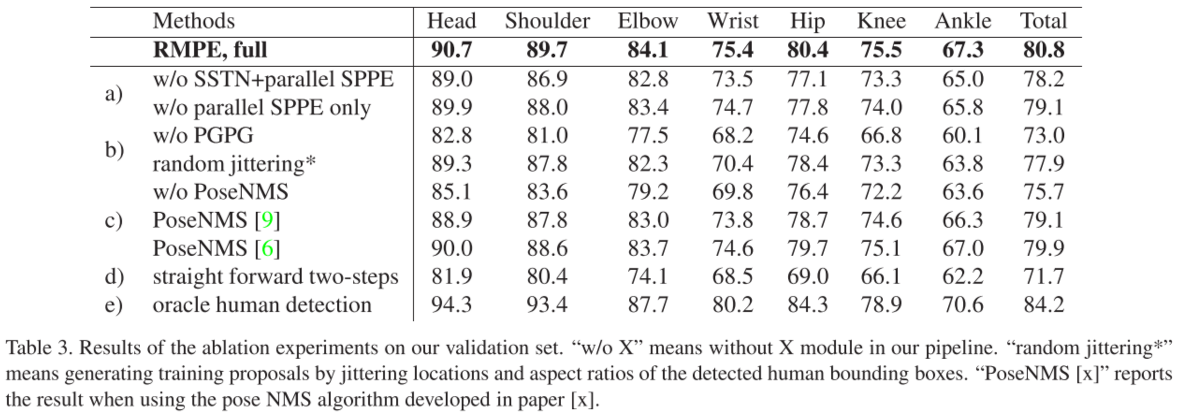
实验结果



Team	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CMU-Pose[7]	61.8	84.9	67.5	57.1	68.2
G-RMI[30]	68.5	87.1	75.5	65.8	73.3
Mask R-CNN[18]	63.1	87.3	68.7	57.8	71.4
Megvii[10]	72.1	91.4	80.0	68.7	77.2
ours	61.8	83.7	69.8	58.6	67.6
ours++	**72.3**	89.2	79.1	68.0	**78.6**

Table 2. Results on the MSCOCO Keypoint Challenge (AP) dataset [2]. The MSCOCO website provides a technical overview only. Our result is obtained without ensembling. “++” denotes using faster-rcnn with softnms [5] as human detector, PyraNet [45] with input size 320x256 as pose estimator. We only compare to single model results.

Ablation Study



	Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
	RMPE, full	**90.7**	**89.7**	**84.1**	**75.4**	**80.4**	**75.5**	**67.3**	**80.8**
a)	w/o SSTN+parallel SPPE	89.0	86.9	82.8	73.5	77.1	73.3	65.0	78.2
	w/o parallel SPPE only	89.9	88.0	83.4	74.7	77.8	74.0	65.8	79.1
b)	w/o PGPG	82.8	81.0	77.5	68.2	74.6	66.8	60.1	73.0
	random jittering*	89.3	87.8	82.3	70.4	78.4	73.3	63.8	77.9
c)	w/o PoseNMS	85.1	83.6	79.2	69.8	76.4	72.2	63.6	75.7
	PoseNMS [9]	88.9	87.8	83.0	73.8	78.7	74.6	66.3	79.1
	PoseNMS [6]	90.0	88.6	83.7	74.6	79.7	75.1	67.0	79.9
d)	straight forward two-steps	81.9	80.4	74.1	68.5	69.0	66.1	62.2	71.7
e)	oracle human detection	94.3	93.4	87.7	80.2	84.3	78.9	70.6	84.2

Table 3. Results of the ablation experiments on our validation set. "w/o X" means without X module in our pipeline. "random jittering*" means generating training proposals by jittering locations and aspect ratios of the detected human bounding boxes. "PoseNMS [x]" reports the result when using the pose NMS algorithm developed in paper [x].

- 其中random jittering是一种一般的数据增强方法
- oracle human detection是直接利用ground truch bbox来训练模型的结果

问题

什么是`atomic pose`