ICPR_2020

2020/12/10

ICPR 2020
25th INTERNATIONAL CONFERENCE
ON PATTERN RECOGNITION
ONLINE 10 | 15 January 2021

A Multi-Task Neural Network for Action Recognition with 3D Key-Points

Rongxiao Tang*, Luyang Wang*, Zhenhua Guo†
Tsinghua Shenzhen International Graduate School

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
ONLINE 10 | 15 January 2021

01 Abstract

- We propose a novel stereo inspired neural network to generate high quality 3D labels for in-the-wild images. We also devise a geometric searching scheme to further refine the 3D joints.
- We build a large-scale Unity dataset for 2D-to-3D pose estimation training task. Unity dataset covers various actions in our daily life.
- Our multi-task framework for 3D human pose estimation performs favorably against baseline approaches, both quantitatively and qualitatively. Experimental results demonstrate that the proposed dataset can significantly boost the generalization performance on realistic scenes.
- We devise a 3D key-point based method for action classification. Experimental results demonstrate that this method achieves better performance on the Penn and NTU datasets than the baseline model.

02 Introduction

□ Limitations of existing models.

1. Traditional 3D pose estimation can be mainly divided into synthetic image, additional annotation, and 2D-to-3D pose estimation. Both of them have limited details and variety levels of synthetic images or require a large number of manual annotations.

➢ we propose a 3D label generator to automatically generate high quality 3D labels for in-the-wild images.

2. Action recognition based on 2D key points, lack of depth information, can not distinguish similar actions well.

➢ Considering close relationship between human body posture and action classification, we propose an action recognition algorithm based on human body 3D key-points, identifying an action by using the key points of human body in real scene

3. 3D action recognition is primarily based on skeleton information acquired by the depth sensor. They are limited to datasets that provide precise key-points, and eliminating effects of imperfect key-points is a major part of their work.

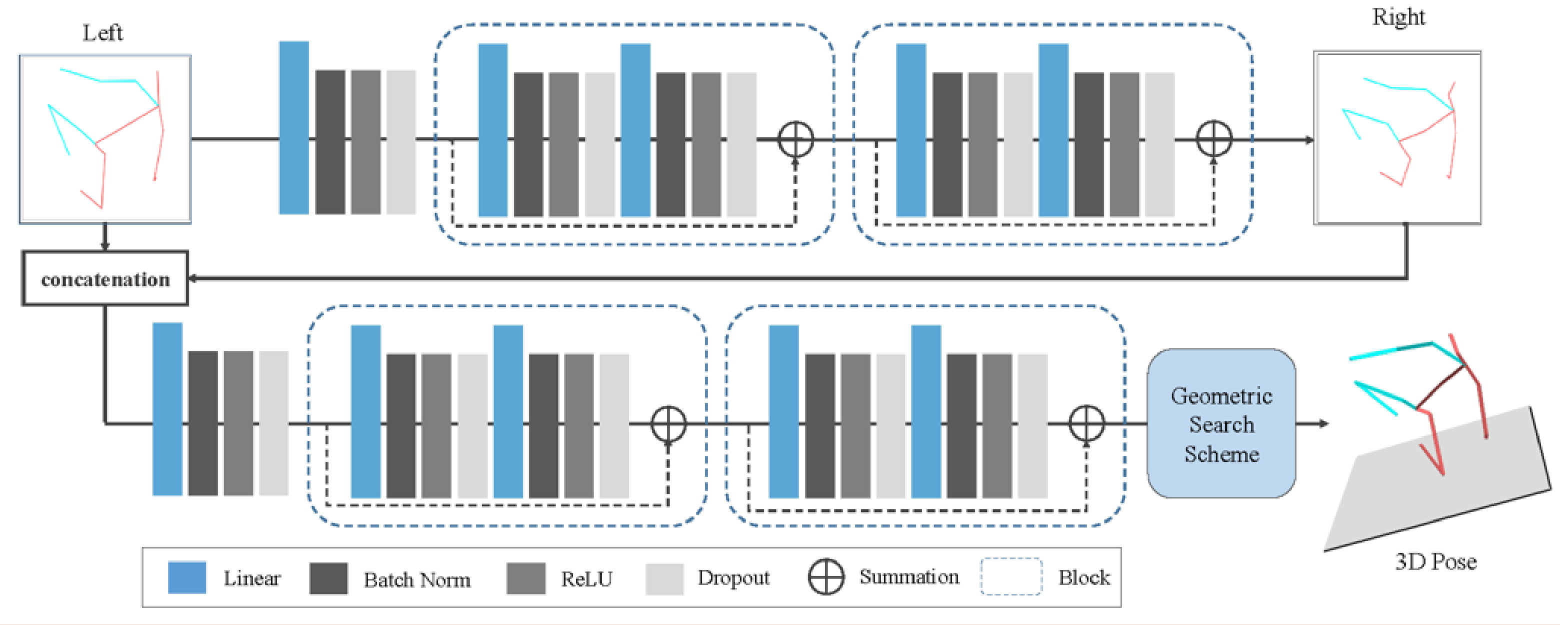
➢ Our method can accurately predict 3D skeleton of a human body based on one input RGB image, so there is no need to deal with noisy human skeletons, which can guide the network for more precise action recognition.

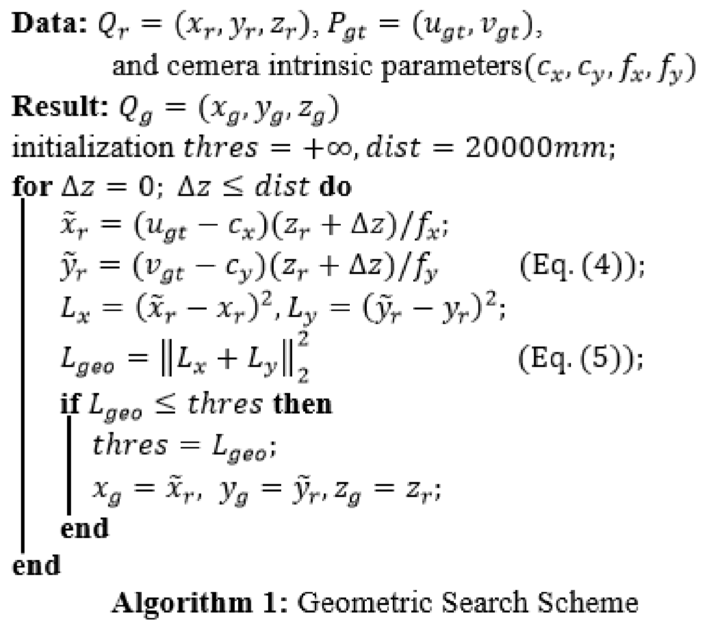
03 Method

➢ 3D Label Generator

- The 3D label generator consists of three parts:

1. Stereoscopic view synthesis subnetwork is proposed to synthesize 2D poses from the right viewpoint.
2. 3D pose reconstruction subnet-work regresses locations of 3D key-points based on the left and right view 2D pose.
3. The geometric search scheme aims to further refine the coarse 3D human pose.





```
Data: $Q_r = (x_r, y_r, z_r)$, $P_{gt} = (u_{gt}, v_{gt})$,
        and cemera intrinsic parameters$(c_x, c_y, f_x, f_y)$
Result: $Q_g = (x_g, y_g, z_g)$
initialization $thres = +\infty, dist = 20000mm$;
for $\Delta z = 0;\ \Delta z \leq dist$ do
    $\tilde{x}_r = (u_{gt} - c_x)(z_r + \Delta z)/f_x$;
    $\tilde{y}_r = (v_{gt} - c_y)(z_r + \Delta z)/f_y$        (Eq. (4));
    $L_x = (\tilde{x}_r - x_r)^2, L_y = (\tilde{y}_r - y_r)^2$;
    $L_{geo} = \left\|L_x + L_y\right\|_2^2$        (Eq. (5));
    if $L_{geo} \leq thres$ then
        $thres = L_{geo}$;
        $x_g = \tilde{x}_r,\ y_g = \tilde{y}_r, z_g = z_r$;
    end
end
```

Algorithm 1: Geometric Search Scheme

Fig.1. Architecture of the 3D label generator.

03 Method

- 3D Action Recognition
- 3D Action Recognition consists of two parts:

1. The 3D label generator is used to generate high quality 3D labels for in-the-wild images or video sequences;
2. The 3D key-points based action recognition neural network consists of Inception V4, 3D hourglass network, and aggregation structure. The aggregation structure is taken from Luvizon [1].

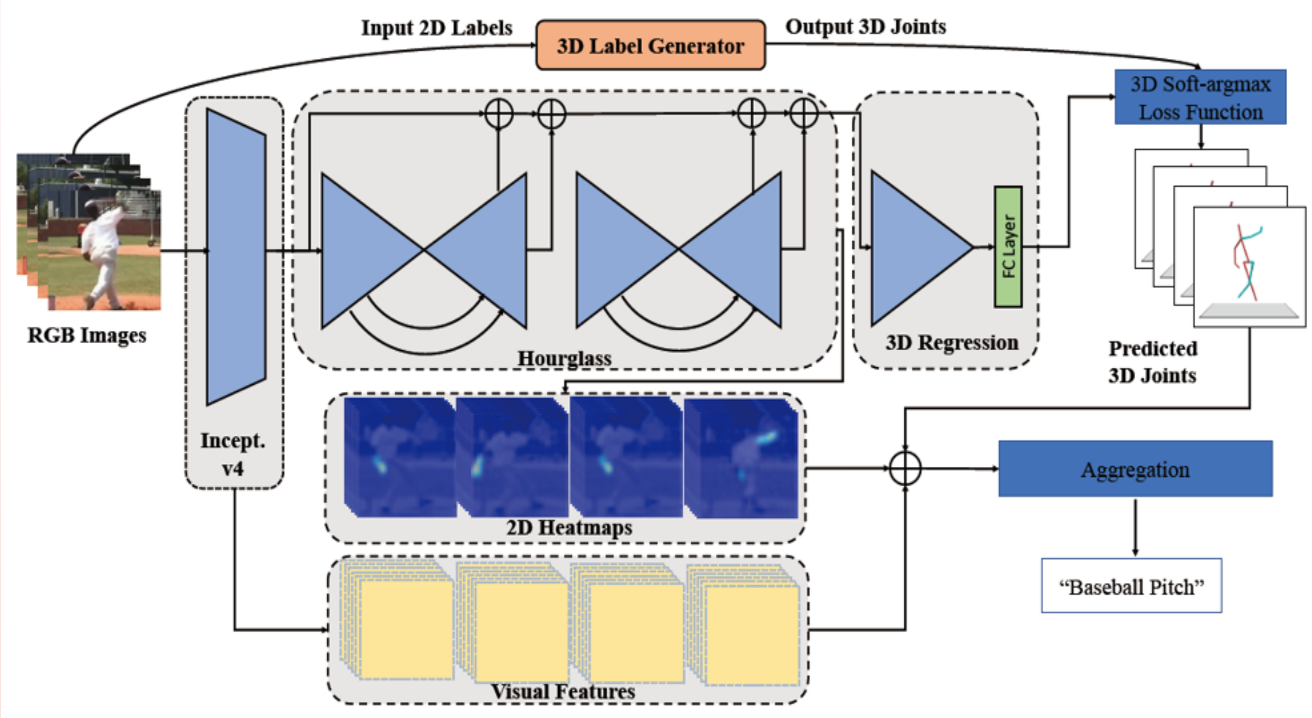


Fig.2. Architecture of the whole pipeline.

[1] Luvizon, Diogo C., David Picard, and Hedi Tabia. "2d/3d pose estimation and action recognition using multitask deep learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

04 Experiments

➢ Implementation Details

1) 3D Label Generator

We train the stereoscopic view synthesis subnetwork with 4.8 million 2D/3D key-points pairs from the Unity toolbox and Human3.6M. When training the 3D pose reconstruction subnetwork, we fix the parameters of the stereoscopic view synthesis subnetwork and adopt the same training scheme.

2)3D Action Recognition

The image and video are processed in the same as Luvizon et al. when training and testing the network. We randomly select fixed-size clips with T frames from a video sample when training, while reporting testing results on single-clip in the middle of the video or multi-clip with temporally spaced of T/2 frame. For the multi-clip, the final score is computed by the average result on all clips from one video.

04

Experiments

➢ Evaluations of 3D Label Generator

1) Quantitative Results

- Table I denotes the comparisons with Martinez et al. [2] on the Human3.6M. All the methods are trained with 2D key-point ground truth. The experimental results show that 3D label generator boosts the performance.
- For protocol#1, the generator trained with 2D/3D ground truth from Human3.6M has 17% (37.6mm vs. 45.5mm) improvements.
- To improve the generalization ability, we also train the network with synthetic 2D/3D pairs generated by the unity toolbox. There is 10% improvement compared with the method of Martinez et al.[2]

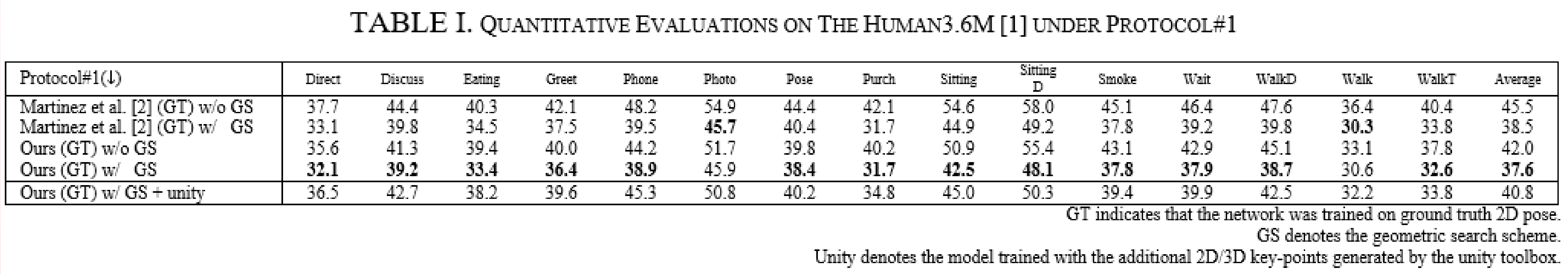


TABLE I. QUANTITATIVE EVALUATIONS ON THE HUMAN3.6M [1] UNDER PROTOCOL#1

Protocol#1(↓)	Direct	Discuss	Eating	Greet	Phone	Photo	Pose	Purch	Sitting	Sitting D	Smoke	Wait	WalkD	Walk	WalkT	Average
Martinez et al. [2] (GT) w/o GS	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
Martinez et al. [2] (GT) w/ GS	33.1	39.8	34.5	37.5	39.5	**45.7**	40.4	31.7	44.9	49.2	37.8	39.2	39.8	**30.3**	33.8	38.5
Ours (GT) w/o GS	35.6	41.3	39.4	40.0	44.2	51.7	39.8	40.2	50.9	55.4	43.1	42.9	45.1	33.1	37.8	42.0
Ours (GT) w/ GS	**32.1**	**39.2**	**33.4**	**36.4**	**38.9**	45.9	**38.4**	**31.7**	**42.5**	**48.1**	**37.8**	**37.9**	**38.7**	30.6	**32.6**	**37.6**
Ours (GT) w/ GS + unity	36.5	42.7	38.2	39.6	45.3	50.8	40.2	34.8	45.0	50.3	39.4	39.9	42.5	32.2	33.8	40.8

GT indicates that the network was trained on ground truth 2D pose.
GS denotes the geometric search scheme.
Unity denotes the model trained with the additional 2D/3D key-points generated by the unity toolbox.

[2] Andriluka, Mykhaylo, et al. "2d human pose estimation: New benchmark and state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014.

04

Experiments

➢ Evaluations of 3D Label Generator

2) Qualitative Results

- We demonstrate the generalization ability qualitatively on the images from MPII and LSP.
- The proposed 3D label generator outperforms the method of Martinez et al. [2]
- The proposed geometric search scheme can refine the coarse 3D human pose.

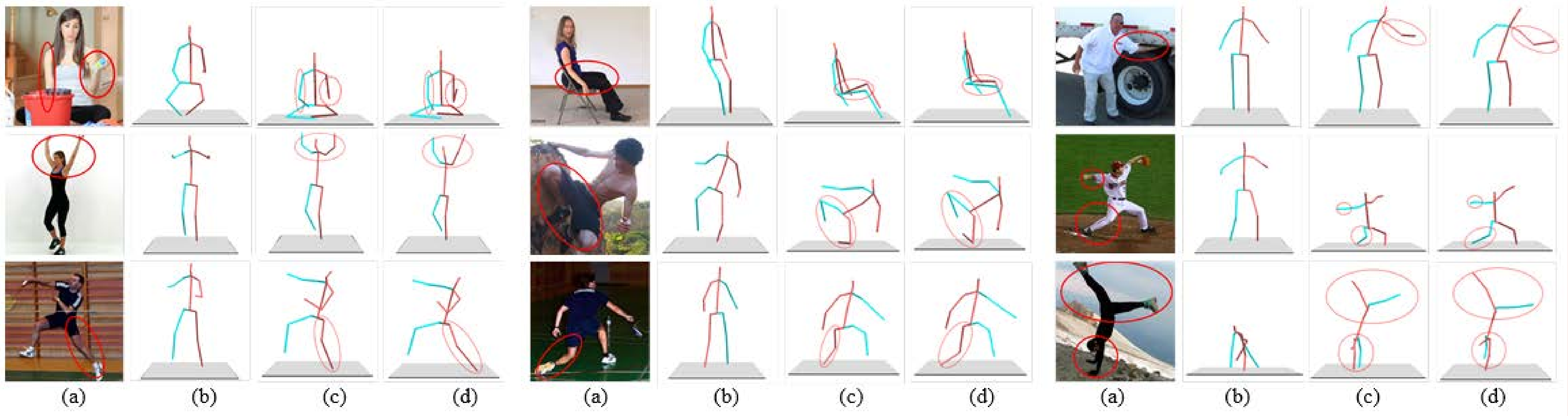


Fig. 3. Qualitative evaluations on in-the-wild images. (a) Original in-the-wild images, (b) Results of Martinez et al. [2], (c) Our results w/o geometric search scheme, (d) Our results w/ geometric search scheme.

[2] Andriluka, Mykhaylo, et al. "2d human pose estimation: New benchmark and state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014.

04 Experiments

➢ Evaluations of 3D Action Recognition Framework

1) 3D action recognition

- We validate it on the Penn and NTU datasets, both of which have valid human 2D/3D key point coordinates.
- The experimental results on the simple Penn dataset are shown in Table II. , we have a 0.7% accuracy improvement.
- The experimental results of the more difficult NTU dataset are shown in Table III. Compared to the best method proposed by Luvizon et al.[1], our method has improved by more than 2% on the NTU dataset.

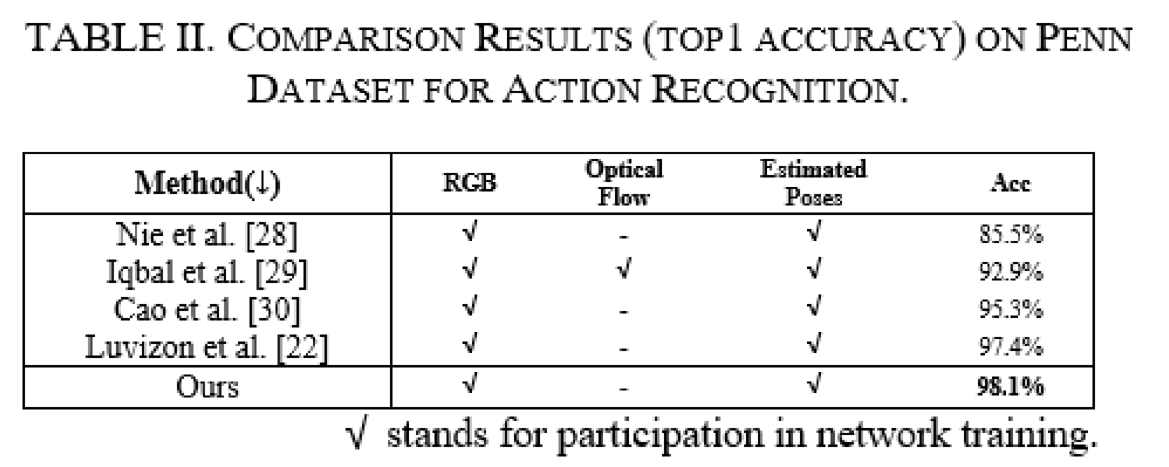


TABLE II. COMPARISON RESULTS (TOP1 ACCURACY) ON PENN DATASET FOR ACTION RECOGNITION.

Method(↓)	RGB	Optical Flow	Estimated Poses	Acc
Nie et al. [28]	√	-	√	85.5%
Iqbal et al. [29]	√	√	√	92.9%
Cao et al. [30]	√	-	√	95.3%
Luvizon et al. [22]	√	-	√	97.4%
Ours	√	-	√	**98.1%**

√ stands for participation in network training.

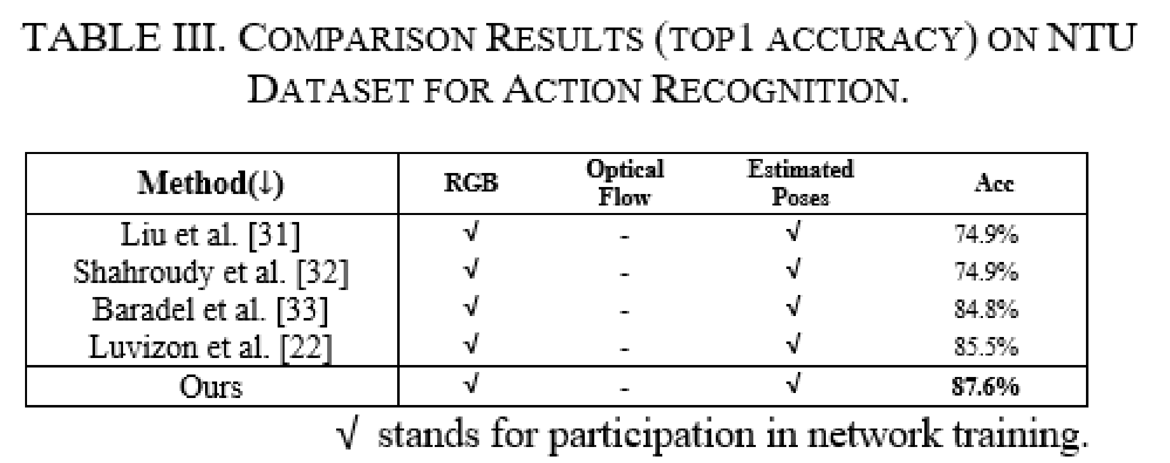


TABLE III. COMPARISON RESULTS (TOP1 ACCURACY) ON NTU DATASET FOR ACTION RECOGNITION.

Method(↓)	RGB	Optical Flow	Estimated Poses	Acc
Liu et al. [31]	√	-	√	74.9%
Shahroudy et al. [32]	√	-	√	74.9%
Baradel et al. [33]	√	-	√	84.8%
Luvizon et al. [22]	√	-	√	85.5%
Ours	√	-	√	**87.6%**

√ stands for participation in network training.

[1] Luvizon, Diogo C., David Picard, and Hedi Tabia. "2d/3d pose estimation and action recognition using multitask deep learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

04 Experiments

➢ Evaluations of 3D Action Recognition Framework

2) Visualization Results

Our multi-task deep learning algorithm can estimate more reasonable results, which in term proves the quality of our proposed dataset, and the effectiveness of our architecture. In addition, our model can handle challenging samples such as leaning over. Experimental results demonstrate that our model has more powerful generalization ability.

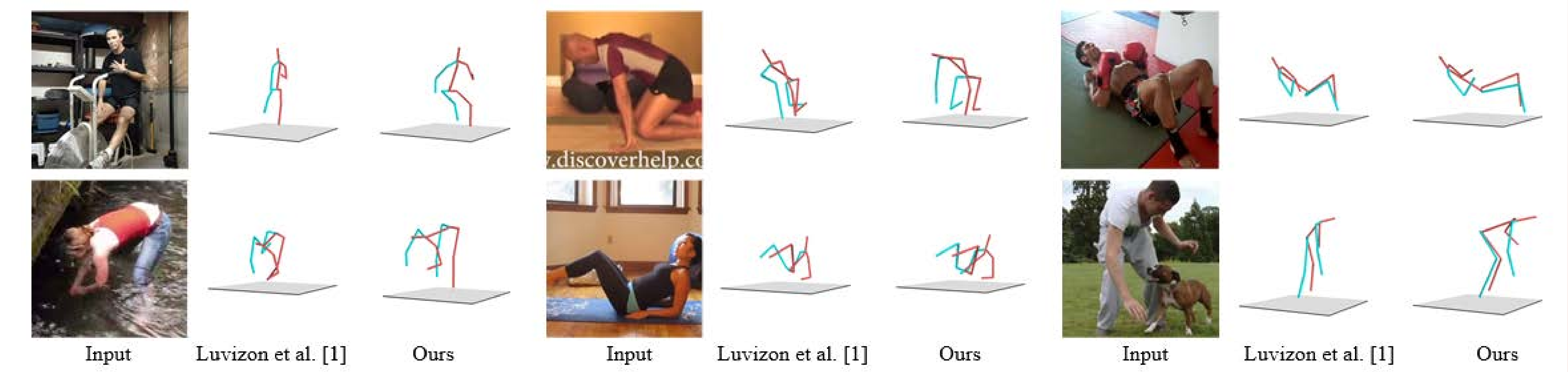


Fig. 4. Qualitative evaluations on in-the-wild images. (a) Original in-the-wild images, (b) Results of Luvizon et al. [1], (c) Our results.

[1] Luvizon, Diogo C., David Picard, and Hedi Tabia. "2d/3d pose estimation and action recognition using multitask deep learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

05 Conclusion

- In this paper, we devise a multi-task neural network for 3D pose estimation and action classification.
- To solve the generalization problem of traditional 3D pose estimation methods, we propose a stereo inspired 3D Label Generator. Based on the stereo inspired structure, the proposed network with a carefully designed geometric search scheme significantly outperforms other methods quantitatively and qualitatively.
- Furthermore, we also devise a 3D joint based deep neural network for action classification. Compared with previous method using merely 2D joints, our method outperforms them on action classification. In addition, our network also utilizes shallow and deep features for action classification. Experimental results demonstrate that our method can achieve better performance on two public datasets, and the 3D pose estimation task can boost performance on action classification.

Thank you!