Category Map Guided Ordinal Depth Prediction for 3D Human Pose Estimation

Liguo jiang
University of the Chinese Academy of Sciences, Institute of Automation

ABSTRACT

In this paper, we propose a two-stage method to estimate 3D human pose, which focuses on the uncertainty of lifting 2D detected pose to 3D pose. Firstly, a novel category map is introduced to predict the ordinal depth category which depicts three kinds of depth ordering relationship for linked joints. Compared with the common probability of vector, our category map can provide better association between prediction with image appearance, and lead to a higher classification accuracy. Secondly, taking predicted 2D pose and ordinal depth category as input, we put forward a temporal convolution network to regress 3D pose, which exploits the temporal context to alleviate the 2D-to-3D uncertainty and reduce prediction errors rate from single image further. Experimental results show that our method can outperform promising results on several benchmarks.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision problems.

KEYWORDS

2D-to-3D pose, ordinal depth category, category map formulation

ACM Reference Format:
Liguo jiang. 2021. Category Map Guided Ordinal Depth Prediction for 3D Human Pose Estimation. In *2021 13th International Conference on Bioinformatics and Biomedical Technology (ICBBT '21), May 21–23, 2021, Xi'an, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3473258.3473303

1 INTRODUCTION

Estimating 3D human pose from monocular images or video is an important problem among the computer vision and graphics fields. Since it provides enormous potential for numerous applications, such as human-computer interaction, video game, medical analytics and social behavior recognition, and so on.

To tackle this problem, many methods make attempts on a two-stage route [1–3]. They usually locate the 2D joint position on the image plane firstly, and then lift it into 3D pose. This route is attractive. On the one hand, it can make full use of the well-studied 2D pose estimation and large images with labeled 2D poses. On the other hand, the abstract 2D joint formulation makes 2D-to-3D regression more focused on human pose itself, rather than complex lighting, background and various human image appearance. Recent work [1] had witnessed the success of 3D pose estimation using only detected 2D joint as input. However, this 2D-to-3D lifting regression is an ill-posed problem inherently, since multiple 3D poses will map a same 2D pose on the image.

To tackle the ambiguity issue, some methods embedded the kinematic priors of human structure into the regression module. For instance, [4] represented human poses as a structure graph encoding, which easily enables the learning of related human joints. In addition, several other methods attempted to exploit extra features to reduce the ambiguity, like implicit image features [5, 23], temporal context in videos [6, 7], pose attributes [2, 34].

In this paper, we attempt to combine the temporal information and part ordinal depth category to estimate 3D pose. Part ordinal depth category encodes rich 3D information, since it measures the three kinds of depth ordering for two linked joints. Compared with the full depth ordering of one joint with respect to all other joints [8], which bears heavy labeling burden, it is easy to make an ordinal depth label for human part. Inspired by semantic segmentation [9], we propose a category map for each part, which records the probability of ordinal depth category on the map. Category map associates strongly part ordering information with 2D appearance in the image. Our experiments witness its superior performance compared with vector-like classification formulation in ImageNet classification [10] and heatmap triplets formulation [11]. Taking the predicted 2D pose and ordinal depth category as input, we adopt the temporal convolution network [7] to regress 3D pose, which exploits the temporal information to alleviate the 2D-to-3D uncertainty and prediction errors from single image further.

The contributions of our methods are summarized as follows:

- We introduce a simple yet effective category map formulation for human part ordinal depth category prediction, which provides a tight association between prediction and image appearance.
- We show that part ordinal depth category helps to resolve the depth ambiguity which cannot be resolved by using temporal context.
- Our method achieves the state-of-the-art performance on both Human3.6M [12] dataset and HumanEva-I dataset [13].

2 RELATED WORKS

Our work focuses on estimating 3D pose from images. And we make a briefly review for two popular routes of 3D pose estimation, one-stage approach and two-stage approach.

2.1 One-stage approach

One-stage approaches usually directly regress 3D pose from images [14, 33, 35]. Pavlakos et al. [14] proposed a volumetric heatmap to

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICBBT '21, May 21–23, 2021, Xi'an, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8965-5/21/05…$15.00
https://doi.org/10.1145/3473258.3473303

depict the joint location probability in space, and used a coarse-to-fine strategy to refine the volumetric heatmaps prediction. However, this volumetric representation usually suffers from heavy storage and computation cost. [15] addressed this issue by introducing a learnable volumetric heatmap autoencoder to make a compression. Popa et al. [16] designed a multi-task network which exploits related tasks (human parsing, 2D pose estimation) to boost the performance of 3D pose prediction. [17] adopted the similar idea as [16], but explored the relation of 3D pose estimation task and human action recognition task.

Above methods heavily rely on large-scale images with ground truth of 3D human poses. Existing 3D datasets, like Human3.6M [12], lacks enough background, lighting and human cloth variations, which brings a big domain gap with wild images. Recent works have attempted to tackle this problem. Zhou et al. [18] proposed a geometric loss which utilizes human structure prior like symmetry, bone length proportion, to restrict wild images. Yang et al. [19] introduced adversarial learning to adapt the domain difference between indoor images and wild images. Sun et al. [20] proposed soft argmax operation which allows 2D/3D mix training uniformly. Although these methods achieve success, but they still easily fall into overfitting, and are sensitive to irrelevant image features.

2.2 Two-stage approach

Two-stage approach is another popular strategy for 3D pose estimation. It decouples the problem into two sub tasks, 2D joint detection [21, 22, 24–26] and 2D-to-3D lifting regression [1, 3, 4]. Martinez et al. [1] designed a simple yet effective network which maps 2D detected joints into 3D joints. Zhao et al. [4] predicted 3D poses from 2D joints through a semantic graph convolutional network, which exploits human topological structure well. Our method also falls into this category.

However, 2D-to-3D lifting is an ill-posed problem inherently. Since many 3D poses share the same mapping of 2D pose in single image. Recently, some works have investigated how to solve the ambiguity problem. Tekin et al. [5] proposed a trainable feature fusion scheme, which effectively fuses the image features and 2D heatmap features to regress 3D joints. Wang et al. [2] predicted additional pose attributes, which depict the relative location of a limb with respect to the torso, and then fed them together with 2D pose into a regression model. Our work also predicts auxiliary cues from images through a multi-task network. Different from [2], we model ordinal depth relation of human parts.

There are also some works have made efforts in 2D-to-3D pose estimation by utilizing temporal context [6, 7, 27]. Hossain et al. [27] introduced a method depicting how to use LSTM units to capture the temporal consistency of 2D pose sequence input, and it produced a sequence of 3D pose which has the same length as input. On the contrary, Pavllo et al. [7] used effective 1D convolutions over 2D pose sequences to capture the temporal context. This model is more effective than [27], due to the inherent parallel processing of convolution operation. Liu [29] et al. added a graph convolution operation to well exploit the local and global spatial information of pose itself, but still used temporal convolution to capture the temporal context. Above methods take only 2D joints sequence as input, and our work proved that it is not sufficient for 3D pose estimation even by utilizing temporal context.

3 METHOD

The overall framework is illustrated in Figure 1. Given a sequence of image frames, a multi-task network is firstly used to predict 2D pose (section 3.1) and part ordinal depth category (section 3.2) frame by frame. And then, a temporal convolution network (section 3.3) takes a series of consecutive predicted information as input, and predicts a 3D human pose. More details are shown in following sections.

3.1 2D pose estimation module

There are many works pay attention to 2D pose estimation [21, 22, 25, 26]. In our works, we adopt a very simple network in [25] as our 2D pose estimator. Given an RGB image $I \in R^{h \times w}$ containing a human subject, we use ResNet-50 [30] to extract image feature $F \in R^{h/32 \times w/32}$, and then followed by three sequential layers (Deconvolution-BatchNorm-ReLU) to increase the resolution of feature maps, finally a 1×1 convolution layer is used to output a set of $\mathcal{J}$ heatmaps. Please note that each deconvolution layer has 256 filters with 4×4 kernel. Heatmap $H_j \in R^{h/4 \times w/4}, j = \{0, 1, \ldots, J\}$ models the probability of one joint j appearing on the map. The whole network structure is shown in Figure 2

Same as [20], we use *soft argmax* loss to train this module, since it can avoid introducing inevitable quantization errors introduced by down sampling operations in convolutional neural network. Let us define the predicted heatmap as H_j for joint j, the predicted 2D coordinate $[x_j, y_j]$, and ground truth 2D $[x_j', y_j']$, the loss L_{2D} is defined as follows:

$$L_{2D} = \sum_{j}^{J} \left(\left| x_j - x_j' \right| + \left| y_j - y_j' \right| \right) \tag{1}$$

$$\left[x_j, y_j \right] = \int_{v} v \cdot softmax \left(H_j \right) \tag{2}$$

where v is the 2D grid coordinates on predicted heatmap.

3.2 Part ordinal depth classification module

Lifting 2D joint to 3D pose is uncertain. The motivation of our method to this issue is to put more auxiliary cues into lifting regressor. For this purpose, we consider adding the part ordinal depth category, which models the depth ordinal relation of one joint with respect to its parent joint. We set part ordinal depth as POD for short.

Figure 3(a) illustrates our skeleton definition. The full skeleton consists of $J = 17$ joints, and it is divided into $P = 14$ skeleton parts. We define B_p as one skeleton part, where $p = \{1, 2, \ldots, P\}$. Each part B_p consists of two associate joints m_p (child joint) and n_p (parent joint). Similar as [11], we define the part ordinal depth information l_p as the depth ordering of m_p and n_p, which is formulated as a tri-state function $\varphi(m_p, n_p)$:

$$l_p = \varphi \left(m_p, n_p \right) = \begin{cases} 0 & \left| z_{m_p} - z_{n_p} \right| \leq \varepsilon \\ 1 & z_{m_p} - z_{n_p} > \varepsilon \\ 2 & z_{m_p} - z_{n_p} < -\varepsilon \end{cases} \tag{3}$$

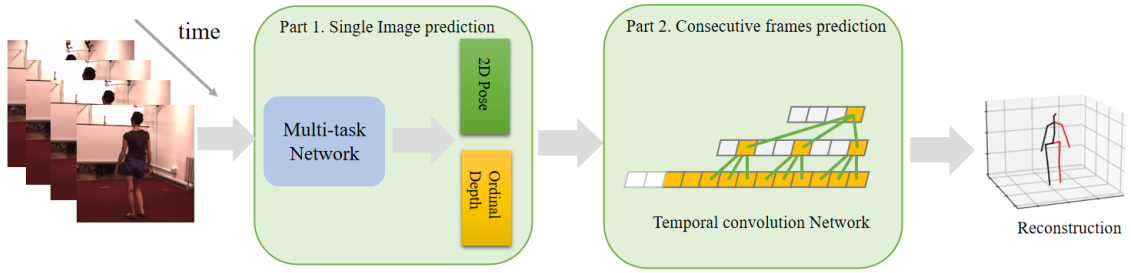

Figure 1: The overview of our proposed method. It mainly consists of two parts: 1) a multi-task network which takes single image as input, and predicts 2D pose and part ordinal depth category simultaneously. 2) a temporal convolution network which takes the predicted 2D pose and part ordinal depth category above from consecutive frames as input, and then predicts 3D human pose.

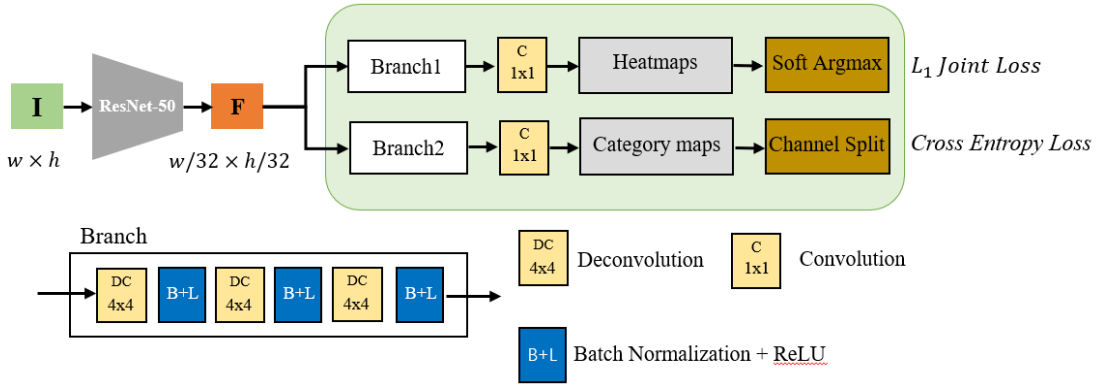

Figure 2: The structure of our multi-task network.

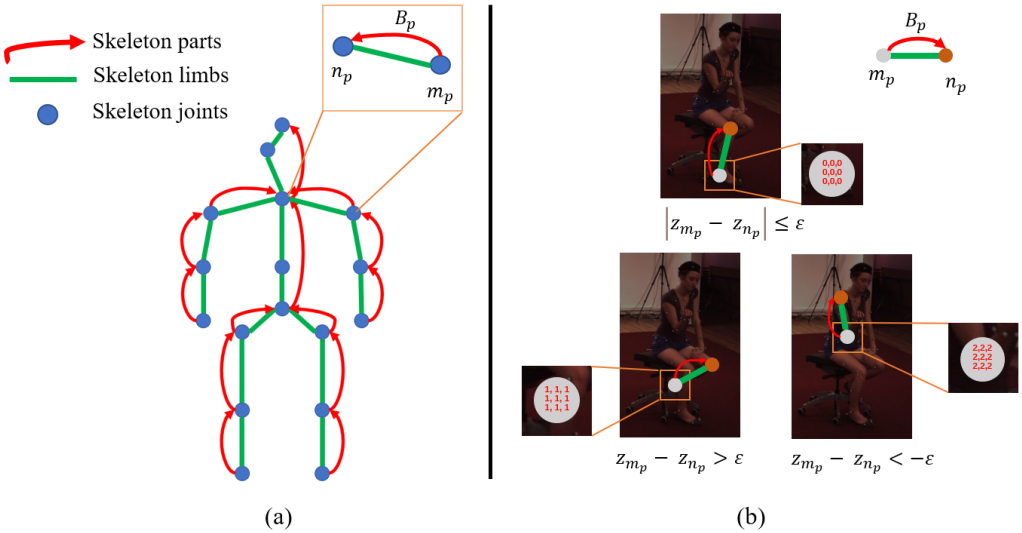

Figure 3: a) Our skeleton consists of 17 joints and 14 parts. b) Tri-state category map configure for parts according to the ordinal depth relationship of two associate joints. Please note that m_p is child joint, n_p is parent joint

where z is the depth value of joint and ε is a sensitive threshold value that controls the labeling sensitivity of depth ordering. We found that setting $\varepsilon = 20mm$ can achieve best result in our experiment. Figure 3(b) shows three kinds of labeling for skeleton parts.

[11] modeled this part depth ordering as heatmap triplets, which was used as intermediate features for predicting 3D pose end-to-end together with image features. On the contrary, we model it as a three-category classification problem for each part directly. Rather than predict a vector-formulated probability for each part, we propose a category map formulation for each part, which was inspired by semantic segmentation. Figure 3(b) shows an example for the ordinal depth category map for B_p, and we set ground truth category on the map around the child joint m_p. Specifically, the predicted part ordinal depth l_p on test phase is set to the category with largest probability at the position of child joint m_p on corresponding predicted probability map. This formulation is linked prediction strongly to image appearance, and our experiments have witnessed its prior performance.

We adopt the same branch structure as above 2D pose estimation module for part depth ordering classification. The network structure is shown in Figure 2. It is essential to split channel by interval 3 at the top of final 1×1 convolutional layer for each part for category loss calculation. Since we are only interested in part depth ordering category of part B_p around child joint m_p, we define the following loss:

$$L_{POD} = \sum_{p}^{P} M_p \odot CrossEntropy\left(O_p, \hat{O}_p\right), \tag{4}$$

where $O_p \in R^{3\times h/4\times w/4}$ is the predicted depth ordering probability map for part B_p, $\hat{O}_p \in R^{h/4\times w/4}$ is the ground truth depth ordering category map, and M_p is the Gaussian mask that indicates the region of child joint m_p nearby.

3.3 Temporal convolutional lifting module

There are two useful units to exploit temporal information, the LSTM units [27] and temporal convolution units [7]. In our work, we use temporal convolution model [7] due to its parallel processing.

The main difference compared to [7] is that we combined detected 2D joint positions and part ordinal category as input. We do not explore the combination ways of these two input features, and just concatenate joint position and part ordinal category in the last axis. And we use $L1$ loss to train this temporal model.

3.4 Implementation details

We implement our two networks in PyTorch [31]. As for the multi-task network, we first initialize the weights of 2D pose estimation module using 2D dataset, as describe in [25], and then we fine tune the full network on Human3.6M [12]. We train the network with batches of 32 and step-wise decreasing learning rate of 5e-5 for every 40k iterations, and 100k iterations for all. The training data is augmented by scale (0.7, 1.35), rotation ($\pm 45^o$), left-right flipping (50% probability), and color distortions when training. As for the temporal convolutional lifting module, we follow the training procedure in [7] by using predicted labels from our multi-task network as training data.

4 EXPERIMENTS

In this section, we firstly describe the datasets and protocols of benchmark we used, and then evaluate the performance quantitatively. We make ablation studies to analyze the effectiveness of proposed method finally.

4.1 Datasets and evaluation protocols

Human3.6M. Human 3.6M [12] is the most popular 3D human dataset. It contains more than 3.6 million RGB images with corresponding ground truth of 3D poses. About 7 female/male subjects (S1, S5, S6, S7, S8, S9, S11) perform 15 different daily actions, like walking, sitting, eating, waiting and so on, in indoor environment. Following the standard protocols for 3D pose estimation [14] in Human3.6M, we use the first 5 subjects (S1, S5, S6, S7, S8) for training and the last 2 subjects (S9, S11) for testing. We evaluate the performance under two standard evaluation protocols in our experiments. Protocol #1 evaluates the mean per joint position error (MPJPE), while protocol #2 measures the MPJPE after procrustes alignment for 3D pose. Both protocols are measured in millimeter.

HumanEva-I. HumanEva-I [13] is another 3D human dataset captured in indoor environment. Compared with Human3.6M, it contains fewer subjects and fewer actions. We evaluated on "Walking" and "Jogging" actions following [14].

4.2 Results and comparisons

Table 1 shows the quantitative comparison results of our method with other related methods under two protocols. It is obvious that our method achieves competitive performance on both protocols. Please note that methods with (*) use temporal context and the others are single-image based methods. Our method achieves near the state-of-the-art results when using temporal context.

We also test our method on HumanEva-I [13], and the quantitative results are presented in Table 2. Our method obtains prior results over previous approaches.

4.3 Ablation study

Effect of depth ordering information. We perform an ablation study to analyze the effect of part ordinal depth category. We use the native work in [7] as baseline. The quantitative results are shown in Table 1. Our method reduces the average error by 5% in protocol#1 and 11% in protocol#2 when compared with baseline [7]. Figure 4 shows a side-by-side comparison of our method with baseline. It is obvious that depth ordering errors exist with only 2D joints as input, like putting hands on back in Figure 4(c), and our method recovers more accurate 3D pose.

Effect of category map formulation. We also analyze the effect of our ordinal depth category map formulation. We compare our map formulation with other two formulations, vector-formulation and heatmap triplets [11]. For vector-formulation, we add a fully-connected layer on image feature F after a global average pooling, to predict a probability vector $v \in R^{42}$, and split it by interval 3 for loss computation. For heatmap triplets, we follow the configure as [11] and use a up sampling branch like ours to predict the heatmap triplets. We adopt the same fine tune procedure as ours on Human3.6M [12] for both baseline training. Furthermore, we evaluate the accuracy on classification of part ordinal depth and

Table 1: Quantitative comparisons on Human3.6M [12] under Protocol #1 and Protocol #2. Best in bold.

Proto#1	Dir	Disc	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WaklD	Walk	WalkT	Avg.
[14]	67.4	71.9	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
[18]	54.8	60.7	58.2	71.4	62.0	65.5	53.8	55.6	75.2	111.6	64.1	66.0	63.2	51.4	55.3	64.9
[1]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
[32]	48.5	54.4	54.4	52.0	59.4	65.3	49.9	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
[5]	44.7	48.9	47.0	49.0	56.4	67.7	48.7	47.0	63.0	78.1	51.1	50.1	54.5	40.1	43.0	52.6
[7]*	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
[6]*	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
[28]*	**41.4**	**44.0**	41.6	**42.6**	46.4	53.4	**41.7**	**41.3**	53.6	60.4	45.8	**41.7**	**45.6**	**32.2**	33.6	**44.3**
Ours*	46.7	44.5	**40.9**	44.2	**46.1**	**52.0**	42.0	41.8	**51.2**	**56.4**	**45.1**	43.4	46.5	33.0	**32.2**	44.4

Protocol #1: Reconstruction errors

Proto#2	Dir	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WaklD.	Walk	WalkT.	Avg.
[14]	47.5	50.5	48.3	49.3	50.7	55.2	46.1	48.0	61.1	78.1	51.1	48.3	52.9	41.5	46.4	51.9
[18]	42.1	44.3	45.0	45.4	51.5	53.0	43.2	41.3	59.3	73.3	51.0	44.0	48.0	38.3	44.8	48.3
[1]	39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
[32]	34.7	39.8	41.8	38.6	42.5	47.5	38.0	36.6	50.7	56.8	42.6	39.6	43.9	32.1	36.5	41.8
[5]	33.6	38.1	37.6	38.5	43.4	48.8	36.0	35.7	51.1	63.1	41.0	38.6	40.9	30.3	34.1	40.7
[7]*	35.7	37.8	36.9	40.7	39.6	45.2	37.4	34.5	46.9	50.1	40.5	36.1	41.0	29.6	33.2	39.0
[6]*	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	**27.3**	36.5
[28]*	**32.1**	35.0	33.5	34.9	36.3	40.9	32.2	**31.8**	42.4	49.0	37.1	**32.4**	**35.6**	**25.0**	27.4	35.0
Ours*	34.2	**33.6**	**33.2**	**34.3**	**34.7**	**40.6**	**32.1**	33.6	**41.3**	**45.9**	**35.3**	32.5	36.1	27.2	27.6	**34.8**

Protocol #2: Reconstruction errors after procrustes alignment

Table 2: Reconstruction errors on the HumanEva-I [13] dataset. All numbers are measured in mm.

Method	Walking			Jogging			Avg
	S1	S2	S3	S1	S2	S3	
[14]	22.3	19.5	29.7	28.9	21.9	23.8	24.3
[2]	18.8	12.7	**29.2**	23.5	15.4	14.5	18.3
[3]	13.9	10.2	46.6	20.9	13.1	13.8	-
ours	**13.1**	**10.1**	45.7	**19.8**	**12.8**	**13.2**	-

Table 3: Quantitively result of three different formulations of part ordinal depth on Human3.6M

Representation	Accuracy of POD (raw) ↑	Accuracy of POD (final) ↑	MPJPE ↓
Vector formulation	76.3	78.5	58.6
Heatmap triplets [11]	83.7	85.0	47.9
Category map (Ours)	86.2	87.6	44.4

mean per joint position error for 3D pose. The quantitative results are shown in Table 3. We set part ordinal depth as POD for short. The accuracy of POD (raw) indicates the classification accuracy by single images as input, while accuracy of POD (final) indicates the classification accuracy of temporal model which trained with the 2D results from images by individual formulation. We can see that our ordinal depth category map formulation works greater than vector formulation and heatmap triplets for both accuracy of POD and MPJPE. In addition, the accuracy POD has increased after temporal model processing, which indicates temporal convolution model helps to reduce the noise of part ordinal classification from single images.

5 CONCLUSIONS

In this work, we present a two-state method for 3D pose estimation. In the first stage, we adopt a multi-task network to predict 2D pose and three kind of part ordinal depth category. The part ordinal depth category, unlike the vector formulation, we model it as a category map formulation which associates the prediction with image appearance tightly. In the second stage, we use a temporal model as 2D-to-3D regression which to alleviate the 2D prediction error

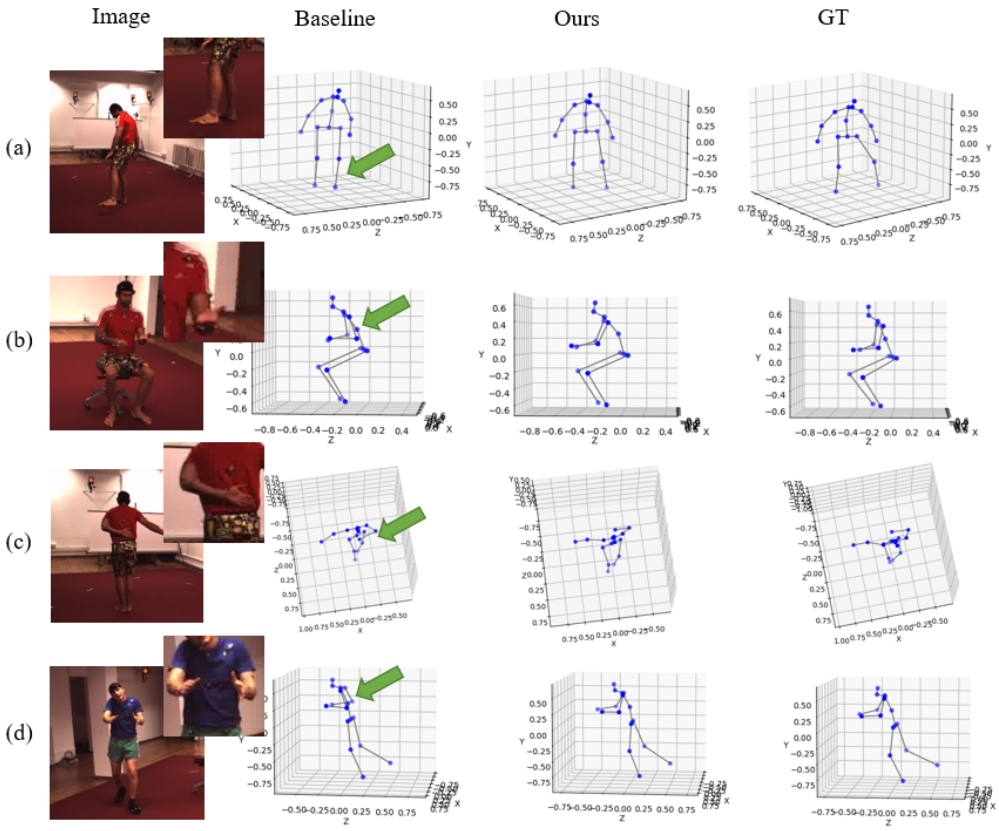

Figure 4: Side-by-side comparisons of our method with baseline [7] on Human3.6M dataset.

and depth ambiguity further. We demonstrate the effectiveness of our method on alleviating depth ambiguity, and achieve state-of-the-art results on benchmarks. In the future, we would like to study the feature combination of 2D joints and the part ordinal depth category, and reduce the impact of part ordinal depth prediction errors. How to improve the generalization performance for wild images is another question we concern.

ACKNOWLEDGMENTS

This research work is supported by The National Science Foundation of China (#U1909218)

REFERENCES

[1] Martinez J, Hossain R, Romero J, *et al.* A simple yet effective baseline for 3d human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2640-2649.
[2] Wang J, Huang S, Wang X, *et al.* Not All Parts Are Created Equal: 3D Pose Estimation by Modeling Bi-directional Dependencies of Body Parts[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 7771-7780.
[3] Fang H S, Xu Y, Wang W, *et al.* Lea3rning pose grammar to encode human body configuration for 3d pose estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
[4] Zhao L, Peng X, Tian Y, *et al.* Semantic graph convolutional networks for 3D human pose regression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3425-3435
[5] Tekin B, Márquez-Neila P, Salzmann M, *et al.* Learning to fuse 2d and 3d image cues for monocular body pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3941-3950.
[6] Cai Y, Ge L, Liu J, *et al.* Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 2272-2281.
[7] Pavllo D, Feichtenhofer C, Grangier D, *et al.* 3d human pose estimation in video with temporal convolutions and semi-supervised training[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7753-7762.
[8] Wang M, Chen X, Liu W, *et al.* Drpose3d: Depth ranking in 3d human pose estimation[J]. arXiv preprint arXiv:1805.08973, 2018.
[9] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
[10] Deng J, Dong W, Socher R, *et al.* Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
[11] Zhou K, Han X, Jiang N, *et al.* HEMlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 2344-2353.
[12] Ionescu C, Papava D, Olaru V, *et al.* Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(7): 1325-1339.
[13] Sigal L, Balan A O, Black M J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion[J]. International journal of computer vision, 2010, 87(1-2): 4.
[14] Pavlakos G, Zhou X, Derpanis K G, *et al.* Coarse-to-fine volumetric prediction for single-image 3D human pose[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7025-7034.
[15] Fabbri M, Lanzi F, Calderara S, *et al.* Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 7204-7213.
[16] Popa A I, Zanfir M, Sminchisescu C. Deep multitask architecture for integrated 2d and 3d human sensing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6289-6298.
[17] Luvizon D, Picard D, Tabia H. Multi-task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[18] Zhou X, Huang Q, Sun X, *et al.* Towards 3d human pose estimation in the wild: a weakly-supervised approach[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 398-407.
[19] Yang W, Ouyang W, Wang X, *et al.* 3d human pose estimation in the wild by adversarial learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5255-5264.
[20] Sun X, Xiao B, Wei F, *et al.* Integral human pose regression[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 529-545.
[21] Wei S E, Ramakrishna V, Kanade T, *et al.* Convolutional pose machines[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.
[22] Papandreou G, Zhu T, Kanazawa N, *et al.* Towards accurate multi-person pose estimation in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4903-4911.
[23] Habibie I, Xu W, Mehta D, *et al.* In the wild human pose estimation using explicit 2d features and intermediate 3d representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10905-10914.
[24] Yang J, Liu Q, Zhang K. Stacked hourglass network for robust facial landmark localisation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 79-87.
[25] Andriluka M, Pishchulin L, Gehler P, *et al.* 2d human pose estimation: New benchmark and state of the art analysis[C]//Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014: 3686-3693.
[26] Chen Y, Wang Z, Peng Y, *et al.* Cascaded pyramid network for multi-person pose estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7103-7112.
[27] Rayat Imtiaz Hossain M, Little J J. Exploiting temporal information for 3d human pose estimation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 68-84.
[28] Chen T, Fang C, Shen X, *et al.* Anatomy-aware 3D Human Pose Estimation in Videos[J]. arXiv preprint arXiv:2002.10322, 2020.
[29] Liu J, Guang Y, Rojas J. GAST-Net: Graph Attention Spatio-temporal Convolutional Networks for 3D Human Pose Estimation in Video[J]. arXiv preprint arXiv:2003.14179, 2020.
[30] He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[31] Paszke A, Gross S, Massa F, *et al.* Pytorch: An imperative style, high-performance deep learning library[C]//Advances in neural information processing systems. 2019: 8026-8037.
[32] Pavlakos G, Zhou X, Daniilidis K. Ordinal depth supervision for 3d human pose estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7307-7316.
[33] Mehta D, Sridhar S, Sotnychenko O, *et al.* Vnect: Real-time 3d human pose estimation with a single rgb camera[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-14.
[34] Ronchi M R, Mac Aodha O, Eng R, *et al.* It's all Relative: Monocular 3D Human Pose Estimation from Weakly Supervised Data[J]. arXiv preprint arXiv:1805.06880, 2018.
[35] Sun X, Shang J, Liang S, *et al.* Compositional human pose regression[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2602-2611.