Patch-based 3D Human Pose Refinement

Qingfu Wan
Fudan University
qfwan13@fudan.edu.cn

Weichao Qiu
Johns Hopkins University
qiuwch@gmail.com

Alan L. Yuille
Johns Hopkins University
alan.l.yuille@gmail.com

Abstract

State-of-the-art 3D human pose estimation approaches typically estimate pose from the entire RGB image in a single forward run. In this paper, we develop a post-processing step to refine 3D human pose estimation from body part patches. Using local patches as input has two advantages. First, the fine details around body parts are zoomed in to high resolution for preciser 3D pose prediction. Second, it enables the part appearance to be shared between poses to benefit rare poses. In order to acquire informative representation of patches, we explore different input modalities and validate the superiority of fusing predicted segmentation with RGB. We show that our method consistently boosts the accuracy of state-of-the-art 3D human pose methods.

1. Introduction

The problem of 3D human pose estimation, defined as localizing 3D semantic keypoints of the human body, has enjoyed substantial progress in recent years [35][34][48][31][22][32]. However, the prediction on some cases are still not accurate enough, especially on poses rarely seen in the training set (*rare* poses). This is due, in large part to the dataset imbalance. Data-driven methods trained on dataset with frequently seen poses (*common* poses) cannot generalize well to *rare* poses [17]. The imbalance between poses makes training difficult, which leads to a model that cannot generate sufficiently accurate result.

To improve 3D human pose estimation, this paper aims at using high-resolution patches that are cropped based on 2D keypoints. Body part patches can produce more accurate result for two reasons. First, computational resource can be gathered to focus on a high-resolution local region. Existing human pose estimation methods usually resize the input image to a fixed scale, in which some body parts have low resolution (See Fig. 2). The fine details in parts are therefore downplayed. To recover high resolution from low resolution, we select the "zoom in" operation which is widely used in lots of vision tasks *e.g.* human part segmentation [45]. Second, the local patch appearance can be shared among different poses. For instance, consider the *rare* sitting pose and *common* standing pose in Fig. 3, their local image appearance around *left_knee* $\rightarrow$ *left_ankle* are similar despite the varied global image appearance. This enables us to train the model via patches from different poses.

In this work, we propose a patch-based refinement module to correct the initial pose estimate of an existing method. Our method upsamples individual local body part patches as input to the refinement module. The refinement module then explicitly concentrates on per-part appearance details to generate a more accurate pose estimate. Fig. 1 shows a brief sketch of the pipeline. The articulation of refinement module is motivated by the fact that estimating pose from body part patches alone is difficult without the global context and skeleton structure constraint. The holistic reasoning of the pose, in fact, conveys valuable information *e.g.* joint angle limit. For this reason, instead of directly estimating 3D pose from patches, we design a refinement module that uses estimation of existing method as an initialization.

To further strengthen the representation of local patches, we use predicted segmentation along with RGB. Predicted segmentation provides useful shape prior for estimating relative depth, while being robust to dim illumination and cluttered background. In occlusion cases, predicted segmentation preserves the occlusion relationship between occluding and occluded body parts.

Our patch-based refinement module can be appended to any existing method. Extensive experiments confirm that the refinement module can effectively improve various state-of-the-art methods. To the best of our knowledge, this is the first successful attempt at improving 3D pose accuracy by patch-based refinement. The refinement is widely applicable with minimal time overhead.

We make the following contributions:

- For the first time, we show that patch-based refinement is able to improve the accuracy of existing 3D human pose methods.
- We demonstrate that high-resolution local part patches

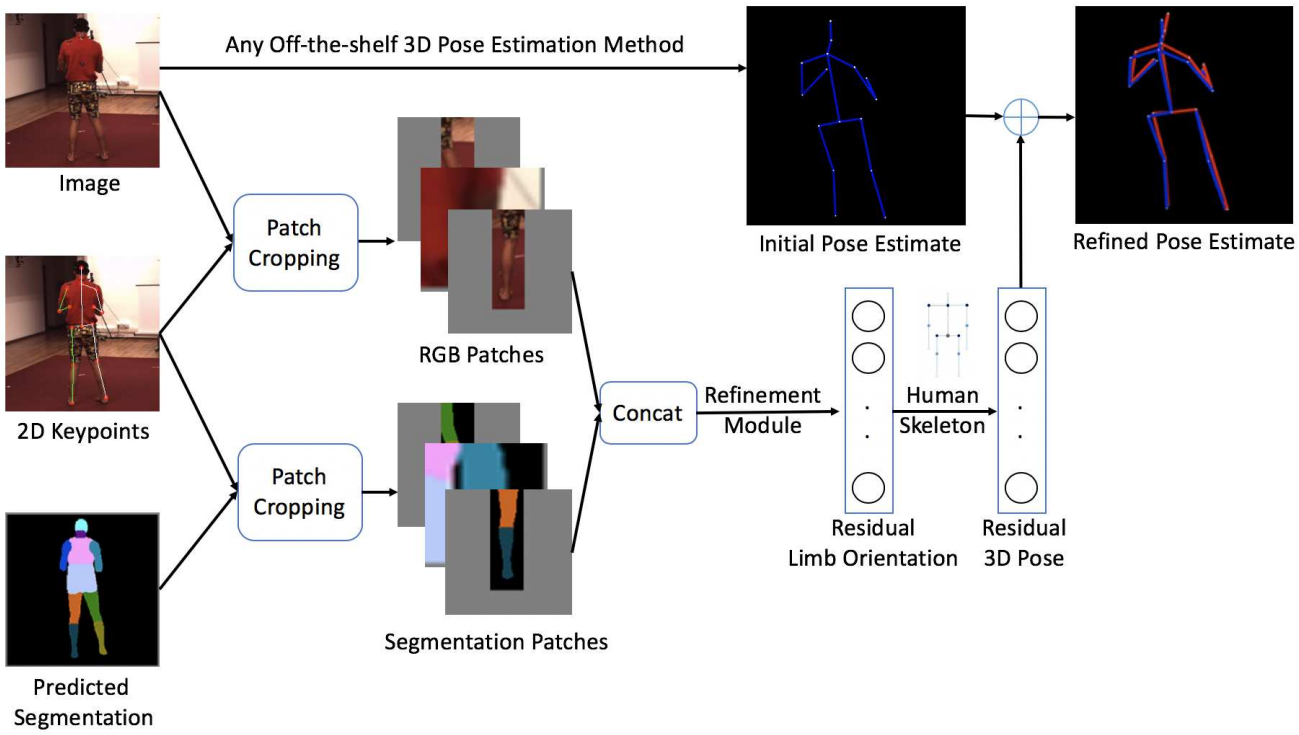

Figure 1. **Framework overview.** Starting from RGB image input, 2D keypoints and segmentation are predicted first. Predicted 2D keypoints are used to crop patches from both RGB and predicted segmentation (color encoded). The RGB patches and segmentation patches are fused together to attain residual limb orientation vector (in 3D), which is transformed to residual 3D pose along the hierarchical human skeleton tree. Residual and initial 3D pose estimate are combined to construct the final 3D estimate. The refined pose (Red) is overlaid on the initial pose (Blue) for better readability. Poses are visualized in a novel 3D viewpoint.

retain fine details to achieve more accurate 3D human pose prediction, especially on *rare* poses.

- We show refinement solely with RGB patches surpasses the original result. Furthermore, we consolidate the extra value of predicted segmentation patches.

2. Related Works

3D human pose estimation 3D human pose estimation has basically been approached in two ways. The first way is to decompose the problem into two steps where the first step estimates 2D from RGB, and the second step lifts 2D to 3D. [22] demonstrate very promising result with a simple multilayer perceptron using 2D skeletal joints as the only input. In similar work, [43] propose to estimate relative depth from skeleton label map[46]. More recently, [29] explore different input representations and establish a very solid system using color-encoded segmentation alone. The performance of these methods is limited, though, owing to the inherent depth ambiguity problem from 2D-3D lifting. [17] argue that generating multiple hypotheses is more reasonable provided this fundamental depth ambiguity nature. We take inspiration from the representation in [29] and merge it with original RGB cue.

Another line of works directly regress 3D from RGB image usually featuring a powerful end-to-end deep learning architecture. The major difference from the previous direction lies in the inclusion of RGB image cue where the image appearance also contributes to the estimation of 3D joints. [23] is the first to employ fully convolutional network in 3D human pose. Later [31] showcase a FCN network with volumetric representation. A recent work [35] power this representation with joint training strategy and a strong ResNet-based architecture. [20] regress a novel representation called orientation map by virtue of fully convolutional network. This method then binds orientation together with each limb region, which better associates image regions and 3D predictions. We draw on the success of this orientation representation and association.

Leveraging local part appearance with global context for inference Combining local part appearance with holistic image has been proven beneficial in 2D pose [27][33][11][15][40]. [6] capture spatial relationship within image patches of different parts with DCNN and graphical model. [39] crop features around coarse 2D keypoint prediction to regress 2D offset. In the scenario of 3D pose,

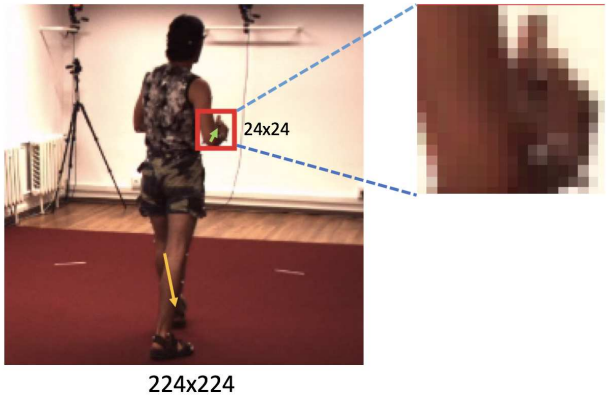

Figure 2. **Motivation: Local patch for fine details amplification.** To determine the orientation of *right_elbow→right_wrist* (Green arrow), we are more interested in the content in Red patch compared to other parts *e.g. left_knee → left_ankle* (Orange arrow). However, the resolution of this "Patch of Interest" (Red) in original 224×224 input image is only 24×24, which is relatively low in modern network architectures. By explicit zoom in operation, the resolution of local patch is increased for further refinement. **Left**: Original image input. **Right**: Recovered high-resolution patch.

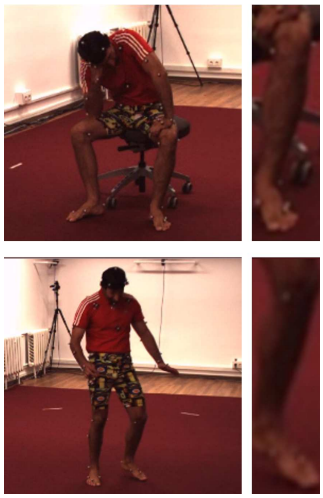

Figure 3. **Motivation: Local patch for appearance sharing.** The local part appearance around *left_knee → left_ankle* is similar between the *rare* sitting pose (**Top**) and the *common* standing pose (**Bottom**), which makes part appearance sharing between *common* and *rare* poses possible. We will later show in Sec 4.4 this is useful for *rare* poses. **Left**: Image input. **Right**: Cropped patch.

however, few works have explicitly processed the information of local part. [5] extract local regional feature map. [25][13][28] perform local 3D refinement from local view in depth image. Different from these works that only utilize depth image as input, we capitalize on local RGB and segmentation cues.

To make use of the low-resolution local image patch, a common strategy is to recover high-resolution map via subsequent upsampling. [45] refine parsing result by adaptively zooming in local region. Lin [19] integrate low-resolution semantic features with fine-grained low-level features to generate high-resolution semantic feature maps. We choose the simple upsampling operation to recover high resolution from low resolution.

Human pose refinement A myriad of methods embed refinement into their pose estimation architectures. [7][26][44][2][3][40] improve 2D keypoint estimation accuracy with multi-stage architecture. [37][38] bring better 3D prediction by repetitive projection and reprojection.

An alternative solution of refinement is to separate the pose estimation and refinement into two parts. Recent work [24] put forward a model-agnostic refinement network by synthesizing pose from error statistics prior. [12] improve the initial estimation by modelling input image space and output pose space. Similarly, our method does not perform pose estimation and refinement in one go.

3. Method

3D human pose estimation targets at localizing predefined 3D keypoints $X \in R^{N\times 3}$ (N is the number of keypoints) from a single RGB image I. Our goal here is to refine the 3D pose output from any existing approach.

The overall architecture is displayed in Fig. 1. To begin with, it takes 3D pose estimation result of any method as initial 3D pose estimate. The patch-based refinement then forwards cropped patches of 2D segmentation and input image to estimate residual pose, which is added with initial pose to output the final refined pose.

3.1. Initial Pose Estimate

Our patch-based 3D pose refinement method is a module that can be attached to any existing 3D pose estimation algorithm. Specifically we deploy existing algorithms [31][30][35][21][18], which take the entire RGB image that encompasses global context as input, to estimate initial 3D prediction $\hat{X}^{(0)}$ from a monocular RGB image.

3.2. Local Patch-based Refinement

Patch Cropping We base the patch cropping operation on 2D keypoint and segmentation prediction. Before cropping patches, we perform 2D keypoint estimation, whereby the keypoints define the local patch region surrounding each body part. We also estimate segmentation S from the input RGB. Note S is a color-coded map from semantic part probability maps.

Having predicted 2D keypoints and segmentation, we crop patches from both RGB and predicted segmentation as follows. For each limb ($N-1$ in total) the predicted

2D of its two endpoints construct a tight bounding box of size $h \times w$, which is center padded and rescaled around part center, so that the patch covers sufficient contextual region. Before feeding into the refinement module, the patch is zero padded and enlarged to network input resolution. This way, the low-resolution patch is zoomed in to offer fine details. The cropping is done on segmentation and RGB respectively. We then concatenate the cropped patches for each limb to form a volume $Concat(Crop(I), Crop(S)) \in R^{((N-1)\times 6)\times H\times W}$, which is the input for the refinement module.

Refinement Module In a nutshell, the objective of refinement is to use local patch details from RGB and segmentation for updating the initial prediction.

$$\hat{X}^{(1)} \leftarrow \hat{X}^{(0)} + Updater(Concat(Crop(I), Crop(S))) \tag{1}$$

Here rather than directly estimate the residual 3D pose $Updater(.)$, we frame the problem as estimating orientation representation introduced in OriNet [20]. Each limb part patch, which is propagated to the refinement module, contains two keypoints attached to that limb. The limb orientation vector represents the relative position between these two keypoints. Thus, the limb orientation representation lends itself natural to model from per-part local appearance. In order to remove the influence of different human scales and resolutions, this orientation vector is additionally normalized by bone length statistic on training set [20]. Since we already have an initialized pose estimate, herein we opt to learn the residual orientation detailed below.

Write $\hat{U}^{(0)}$ as the predicted orientation vector from initial pose estimate $\hat{X}^{(0)}$ in Sec. 3.1 and U^{gt} as the ground truth counterpart, the residual we aim to learn is $U^{gt} - \hat{U}^{(0)}$. We adapt ResNet-50 [14] to learn this residual orientation.

If we denote ΔU as the learnt residual orientation, then the loss function is:

$$\mathcal{L} = \sum_k ||\Delta U_k - (U_k^{gt} - \hat{U}_k^{(0)})||_2^2 \tag{2}$$

where ΔU_k is the learnt residual orientation for the k-th limb.

During inference, the learnt residual orientation ΔU is transformed back to residual 3D pose for final estimation. In more detail, ΔU is scaled back with limb length statistic to $Unnorm(\Delta U)$. We then reconstruct residual 3D pose $Updater(.)$ along the skeleton tree hierarchy with $Unnorm(\Delta U)$, following previous practice [20]. Afterwards we add the residual with initial pose estimate to produce the final refined pose (Eq. (1)).

4. Experiments

4.1. Implementation Details

For 2D keypoints, we apply integral regression [35] on top of keypoint probability maps from 2D Hourglass [26]. For 2D segmentation, we employ NBF [29] for its state-of-the-art accuracy. The part segmentation is color encoded to $3 \times 256 \times 256$. The tight bounding box in Sec. 3.2 is center padded to $max(28, h) \times max(28, w)$. The rescaling factor is empirically set to 2.3. The cropped patches are resized to 256×256 and then fed into a ResNet-50 [14], where the last 1000-way fully connected layer is changed to output 48-D residual orientation vector (Sec. 3.2 ΔU). Weights pretrained on ImageNet [9] are loaded up to the penultimate layer. L2 loss is enforced to learn ΔU (Eq. (2)). We do not perform end-to-end training, but rather take result of other methods as initial pose estimate. Implementations are in Caffe and PyTorch. We train the refinement module for 20 epochs using Adam with batch size of 32. Base learning rate is 1e-5, which is divided by 10 after loss plateau on the validation set.

4.2. Datasets and Metrics

We conduct experiments on Human3.6M [16], which is insofar the largest 3D human pose dataset for indoor MoCap setup. We follow the standard protocol to use subject S1, S5, S6, S7, S8 for training and test on S9, S11 every 64 frames. We measure pose accuracy in terms of MPJPE (mean per joint position error), which has been widely used before [31] [22] [21][48][35].

4.3. Improvement over State-of-the-art Methods

We report the performance improvement when our method is applied to state-of-the-art methods in Tab. 1. We experiment with five methods [31][30][35][21][18]. To obtain initial pose estimate, we use their released code with pretrained models and test by ourselves whenever possible. We can see that the patch-based refinement yields better result, especially for rare poses *e.g. SitDown* on [31][21] and *Sit* on [31][30].

4.4. Qualitative Visualization

To further analyze the improvement, in Fig. 4 we present qualitative result. Two cases, where our local patch-based refinement is of vital importance, are highlighted. As exemplified in Fig. 5 and Fig. 6, almost all the joints are more accurately localized in these two cases.

Fig. 5 shows the first case: *rare pose*. As stated previously, similar local part appearance shared from *common* poses can aid the refinement of *rare* poses.

Fig. 6 visualizes the second case: *occlusion*. When occlusion happens, the additional segmentation cue makes it

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Pavlakos [31]	59.7	70.3	59.0	78.7	64.9	54.7	72.9	80.9
+ **Refinement (Ours)**	**56.5**	**64.4**	**57.5**	**60.1**	**62.5**	**50.9**	**68.9**	**79.4**
Integral Pose [35]	63.3	51.8	54.5	92.4	54.1	45.4	52.7	66.8
+ **Refinement (Ours)**	**60.4**	**51.5**	**54.3**	**80.7**	**53.9**	**45.2**	**52.7**	**66.7**
Luvizon [21]	**55.3**	57.7	51.6	55.9	57.0	**53.5**	56.8	66.8
+ **Refinement (Ours)**	55.4	**57.7**	**51.4**	**55.6**	**56.5**	53.6	**53.3**	**66.0**
Pavlakos *et al*. [30]	47.5	52.6	55.3	50.8	58.5	47.4	52.8	64.5
+ **Refinement (Ours)**	**46.8**	**52.1**	**54.3**	**50.0**	**57.5**	**46.8**	**52.8**	**63.5**
Kocabas [18]	60.9	49.8	46.6	70.1	48.8	45.4	45.6	53.7
+ **Refinement (Ours)**	**60.5**	**49.6**	**46.4**	**70.0**	**48.7**	**45.0**	**45.5**	**53.6**
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Pavlakos [31]	134.6	62.4	78.9	74.6	48.9	69.6	57.0	70.7
+ **Refinement (Ours)**	**120.8**	**59.8**	**76.9**	**57.0**	**45.0**	**66.3**	**54.2**	**65.2**
Integral Pose [35]	104.6	54.6	61.7	68.6	40.9	54.8	46.5	60.9
+ **Refinement (Ours)**	**97.1**	**54.4**	**61.6**	**53.2**	**40.5**	**54.5**	**46.2**	**58.3**
Luvizon [21]	78.3	58.4	65.8	52.5	48.8	62.9	52.0	58.3
+ **Refinement (Ours)**	**77.1**	**58.2**	**65.6**	**52.2**	**48.6**	**62.6**	**51.6**	**57.9**
Pavlakos *et al*. [30]	69.6	54.7	65.2	52.6	44.9	60.0	48.0	55.3
+ **Refinement (Ours)**	**69.6**	**53.9**	**64.2**	**51.8**	**44.3**	**59.1**	**46.8**	**54.5**
Kocabas [18]	87.9	49.2	52.2	46.7	42.6	51.3	45.1	52.8
+ **Refinement (Ours)**	**87.8**	**48.9**	**51.9**	**46.5**	**42.2**	**51.1**	**44.8**	**52.6**

Table 1. **Improvement of MPJPE when the patch-based refinement is applied to state-of-the-art methods.** No procrustes alignment is used. The lower the number, the better the result. Bold face indicates the better result.

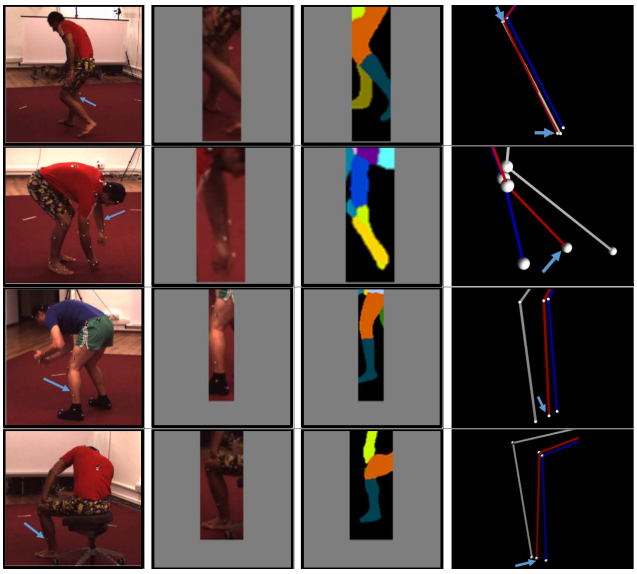

Figure 4. **Qualitative results of the patch-based refinement. Left:** Image input. **Middle:** Cropped segmentation and RGB image patch. **Right:** The refined result (Red) on initial estimate (Blue). Ground truth is colored in white for reference. Blue arrow points to the part and refined joint. Only best 3D local view is visualized.

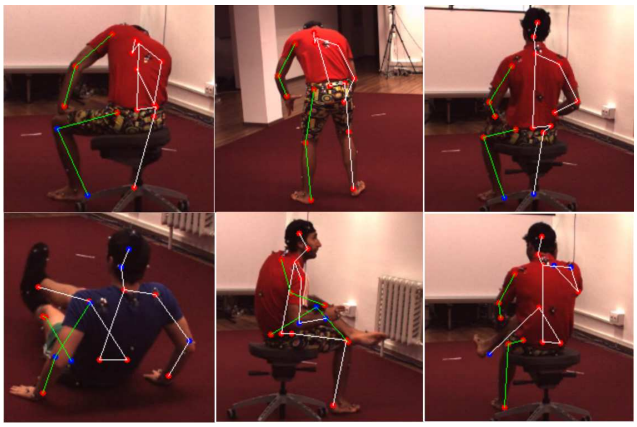

Figure 5. **Most helpful case 1: rare pose**. Red indicates a joint is improved with patch-based refinement. Blue indicates no improvement.

easy to discriminate between occluding and occluded limb. A vivid illustration can be found in Fig. 7.

4.5. Ablation Study

We use the method in Pavlakos [31] to generate initial pose estimate for ablation study. We will first elucidate the importance of patch cropping operation. Then we will

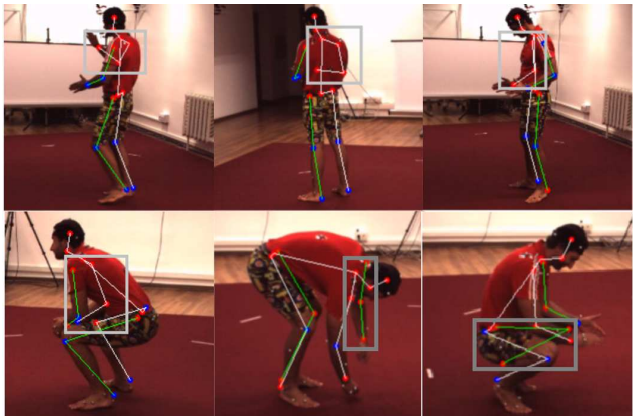

Figure 6. **Most helpful case 2: occlusion**. Red indicates a joint is improved with patch-based refinement. Blue indicates no improvement. Occluded part is enclosed in rectangle.

elaborate on the importance of using both segmentation and RGB for patch cropping.

4.5.1 Importance of Patch Cropping

To prove the necessity of patch cropping operation for refinement, we implement a baseline where the original RGB image rather than the cropped patch is input to the refinement module. No segmentation is used for simplicity. Passing the entire RGB image into the refinement module, which has been explored in [36][44], can be interpreted as stacking one more stage to any prevalent multi-stage pose estimation architecture. As seen in Tab. 2, cropped patch performs generally better than uncropped RGB image. This can be attributed to the high-resolution local patch where local detail is amplified.

4.5.2 Importance of Fusing Segmentation with RGB

Having established that patch cropping is necessary, we now proceed to investigate the best input modality for patch cropping. In Tab. 3, we quantitatively compare different choices of patch input: (1) **w/ cropped RGB**: with only cropped RGB patches. (2) **w/ cropped Seg**: with only cropped segmentation patches. (3) **w/ cropped RGB + cropped Seg**: mixture of cropped segmentation and cropped RGB patches. Among which (3) performs the best. One observation is that (1) is already better than initial pose estimate, which shows the effectiveness of the patch-based refinement. To gain insight on the benefit of the extra segmentation cue, we depict in Fig. 7 two specific cases when using cropped RGB is not accurate enough. In the first case, the cropped RGB patch is too vague to discern among lower arm, upper arm and background. Segmentation gets rid of the background wall and singles out the two arms. The other

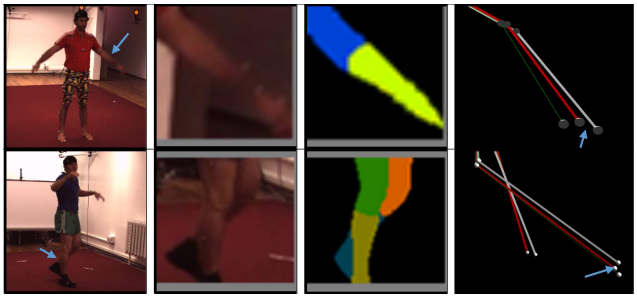

Figure 7. **Qualitative examples showing adding cropped segmentation is better than only cropped RGB. Left**: Image input. **Middle**: Cropped segmentation and RGB image patch. **Right**: The result with cropped segmentation (Red) *vs* with only cropped RGB (Green). White is ground truth. Body part and refined joint are marked with Blue arrow. We only show the novel local 3D view for better readability. **Top**: Note that the lower arm almost blends in with the background, which is eliminated in segmentation. Besides, the dim illumination no longer exists. **Bottom**: Occlusion case. Left and right ankle are not clearly shown in the RGB patch because of the overlapping shoes. In the segmentation patch, nonetheless, left leg and right leg are distinguishable.

case contains an occluded part: *left_knee* $\rightarrow$ *left_ankle*. It is evident that the RGB patch fails to distinguish between left ankle and right ankle, which is addressed by segmentation. When segmentation occasionally fails *e.g.* the shape is completely wrong, the other RGB cue can still prevent the refinement module from outputting a huge residual pose. See [41] for more detailed discussion.

5. Discussion

It should be noted that there are some tricks to further boost the performance. Below we list some examples.

It is feasible to use conditional random field [8], attention mechanism [4] or feature pyramid [47] to further exploit appearance information contained in a local patch. We only consider rescaling all the body part patches to a fixed scale, which is limited in that different body parts may have different sizes. To deal with this issue, parts can be adaptively zoomed in to different proper scales[45]. For simplicity, we here only discuss patch cropping using RGB and segmentation. One can make use of other representations *e.g.* 2D keypoint probability map [36], 2D skeleton label map [46][43], height map [10], star map [49], joint angle [50] *etc*. Our current fully connected regression implementation can also be extended to dense regression by fully convolutional network for preciser prediction [20][42].

As to refinement itself, the current refinement module equally treats joints that are already very close to ground truth and that are far away from ground truth. Confidence-aware refinement can actually be adopted, where individual weights are given to each joint allowing refinement prioriti-

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
+ Refinement (w/ uncropped RGB)	62.0	66.5	58.3	63.0	62.9	57.1	**66.9**	80.7
+ Refinement (w/ cropped RGB)	**57.2**	**65.9**	**58.0**	**61.0**	**62.5**	**52.2**	71.3	**79.3**
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
+ Refinement (w/ uncropped RGB)	**118.2**	61.9	**76.7**	63.1	49.2	**65.9**	56.5	67.2
+ Refinement (w/ cropped RGB)	121.1	**59.9**	77.0	**58.3**	**45.5**	67.2	**54.7**	**66.0**

Table 2. **Necessity of patch cropping operation.** The result of refinement with cropped RGB patches and with original RGB image input on [31]. Segmentation cue is not used here. Using patch is generally better than original RGB image as input for refinement.

Method	Direction	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Pavlakos [31]	59.7	70.3	59.0	78.7	64.9	54.7	72.9	80.9
+ Refinement (w/ cropped RGB)	57.2	65.9	58.0	61.0	62.5	52.2	71.3	79.3
+ Refinement (w/ cropped Seg)	61.5	66.1	58.1	62.3	62.7	55.9	**67.0**	80.5
+ Refinement (w/ cropped RGB + cropped Seg)	**56.5**	**64.4**	**57.5**	**60.1**	**62.5**	**50.9**	68.9	**79.4**
Method	SitDown	Smoke	Photo	Wait	Walk	WalkDog	WalkPair	Avg
Pavlakos [31]	134.6	62.4	78.9	74.6	48.9	69.6	57.0	70.7
+ Refinement (w/ cropped RGB)	121.1	59.9	77.0	58.3	45.5	67.2	54.7	66.0
+ Refinement (w/ cropped Seg)	**117.4**	61.6	**76.0**	61.6	48.5	**65.6**	56.2	66.7
+ Refinement (w/ cropped RGB + cropped Seg)	120.8	**59.8**	76.9	**57.0**	**45.0**	66.3	**54.2**	**65.2**

Table 3. **Effect of different patch input modality.** This table explains the reason to fuse cropped segmentation and cropped RGB.

zation of some joints, in a similar way as [1].

6. Conclusion

We present the first patch-based 3D human pose refinement method. We substantiate that the local body part patches from RGB, which preserve fine details, can be zoomed in to high resolution for accurate prediction. Further, we prove the effectiveness of incorporating segmentation prediction with RGB. We empirically observe that the local part appearance sharing between poses is important for refining *rare* poses. The high-resolution fine details and local appearance sharing result in consistent performance gain on state-of-the-art methods. Our method is model-agnostic, which can be inserted after any 3D pose model to refine inaccurate poses with minimum computational cost.

Acknowledgement

This work is supported by IARPA via DOI/IBC contract No. D17PC00342.

References

[1] D. Antotsiou, G. Garcia-Hernando, and T.-K. Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *ECCV*, 2018. 7

[2] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016. 3

[3] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 3

[4] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016. 6

[5] X. Chen, G. Wang, H. Guo, and C. Zhang. Pose guided structured region ensemble network for cascaded hand pose estimation. *arXiv preprint arXiv:1708.03416*, 2017. 3

[6] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014. 2

[7] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 3

[8] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017. 6

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*. Ieee, 2009. 4

[10] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. Marker-less 3d human motion capture with monocular image sequence and height-maps. In *ECCV*, 2016. 6

[11] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, 2015. 2

[12] M. Fieraru, A. Khoreva, L. Pishchulin, and B. Schiele. Learning to refine human pose estimation. In *CVPR Workshops*, pages 205–214, 2018. 3

[13] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *ICCV*, 2017. 3

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[15] J. Hwang, S. Park, and N. Kwak. Athlete pose estimation by a global-local network. In *CVPR Workshops*, 2017. 2

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 4

[17] E. Jahangiri and A. L. Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *ICCV Workshops*, 2017. 1, 2

[18] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. *arXiv preprint arXiv:1903.02330*, 2019. 3, 4, 5

[19] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 3

[20] C. Luo, X. Chu, and A. Yuille. Orinet: A fully convolutional network for 3d human pose estimation. *arXiv preprint arXiv:1811.04989*, 2018. 2, 4, 6

[21] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018. 3, 4, 5

[22] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 1, 2, 4

[23] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 2

[24] G. Moon, J. Y. Chang, and K. M. Lee. Posefix: Model-agnostic general human pose refinement network. *arXiv preprint arXiv:1812.03595*, 2018. 3

[25] G. Moon, J. Y. Chang, Y. Suh, and K. M. Lee. Holistic planimetric prediction to local volumetric prediction for 3d human pose estimation. *arXiv preprint arXiv:1706.04758*, 2017. 3

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3, 4

[27] X. Nie, J. Feng, J. Xing, S. Xiao, and S. Yan. Hierarchical contextual refinement networks for human pose estimation. *IEEE Transactions on Image Processing*, 28(2):924–936, 2019. 2

[28] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015. 3

[29] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*. IEEE, 2018. 2, 4

[30] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3d human pose estimation. *arXiv preprint arXiv:1805.04095*, 2018. 3, 4, 5

[31] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*. IEEE, 2017. 1, 2, 3, 4, 5, 7

[32] S. Sharma, P. T. Varigonda, P. Bindal, A. Sharma, and A. Jain. Monocular 3d human pose estimation by generation and ordinal ranking. *arXiv preprint arXiv:1904.01324*, 2019. 1

[33] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, and J. Wang. Human pose estimation using global and local normalization. In *ICCV*, 2017. 2

[34] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, 2017. 1

[35] X. Sun, B. Xiao, S. Liang, and Y. Wei. Integral human pose regression. *arXiv preprint arXiv:1711.08229*, 2017. 1, 2, 3, 4, 5

[36] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *ICCV*, 2017. 6

[37] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *CVPR*, 2017. 3

[38] D. Tome, M. Toso, L. Agapito, and C. Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *3DV*. IEEE, 2018. 3

[39] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015. 2

[40] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2, 3

[41] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. *arXiv preprint arXiv:1804.04875*, 2018. 6

[42] C. Wan, T. Probst, L. Van Gool, and A. Yao. Dense 3d regression for hand pose estimation. In *CVPR*, 2018. 6

[43] Q. Wan, W. Zhang, and X. Xue. Deepskeleton: Skeleton map for 3d human pose regression. *arXiv preprint arXiv:1711.10796*, 2017. 2, 6

[44] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 3, 6

[45] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 1, 3, 6

[46] F. Xia, P. Wang, X. Chen, and A. L. Yuille. Joint multi-person pose estimation and semantic part segmentation. In *CVPR*, 2017. 2, 6

[47] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *ICCV*, 2017. 6

[48] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 1, 4

[49] X. Zhou, A. Karpur, L. Luo, and Q. Huang. Starmap for category-agnostic keypoint and viewpoint estimation. In *ECCV*, 2018. 6

[50] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*, 2016. 6