ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation

Yufei Xu[1], **Jing Zhang**[1], **Qiming Zhang**[1], **Dacheng Tao**[2,1]

[1]School of Computer Science, The University of Sydney, Australia

[2]JD Explore Academy, China

Abstract

Recently, customized vision transformers have been adapted for human pose estimation and have achieved superior performance with elaborate structures. However, it is still unclear whether plain vision transformers can facilitate pose estimation. In this paper, we take the first step toward answering the question by employing a plain and non-hierarchical vision transformer together with simple deconvolution decoders termed **ViTPose** for human pose estimation. We demonstrate that a plain vision transformer with MAE pretraining can obtain superior performance after finetuning on human pose estimation datasets. ViTPose has good scalability with respect to model size and flexibility regarding input resolution and token number. Moreover, it can be easily pretrained using the unlabeled pose data without the need for large-scale upstream ImageNet data. Our biggest ViTPose model based on the ViTAE-G backbone with 1 billion parameters obtains the best 80.9 mAP on the MS COCO test-dev set, while the ensemble models further set a new state-of-the-art for human pose estimation, i.e., 81.1 mAP. The source code and models will be released at https://github.com/ViTAE-Transformer/ViTPose.

1 Introduction

Human pose estimation is one of the fundamental tasks in computer vision and has a wide range of real-world applications Zhang & Tao (2020). However, it is very challenging due to the variations of occlusion, truncation, scale, illumination, and human appearances. There has been rapid progress in deep learning-based pose estimation methods Toshev & Szegedy (2014); Xiao et al. (2018); Sun et al. (2019); Zhang et al. (2021) to deal with the difficult task using convolutional based networks.

Recently, inspired by the success of transformer structures in vision tasks Dosovitskiy et al. (2020); Liu et al. (2021); Xu et al. (2021); Zhang et al. (2022a;b), different vision transformer structures have been adapted for pose estimation. With a cascade regression mechanism, PRTR Li et al. (2021a) incorporates an encoder-decoder structure to predict the human keypoints gradually. TokenPose Li et al. (2021b) and TransPose Yang et al. (2021) adopts vision transformer to refine the features extracted by CNN, either with or without additional keypoint tokens. HRFormer Yuan et al. (2021), on the other hand, adapts to the pose estimation tasks by introducing high-resolution representations into vision transformers via multi-resolution parallel transformer modules. These methods have obtained superior performance on human pose estimation tasks. However, it is unclear whether human pose estimation tasks can benefit from the plain vision transformers without these modifications.

In this paper, we take the first step toward answering the question by adopting a plain and non-hierarchical vision transformer for the human pose estimation task, i.e., ViTPose. Specially, we explore the usage of vision transformers for human pose estimation by simply appending several deconvolution layers after the vision transformers to regress the keypoint heatmaps. Without whistles and bells, such a simple baseline obtains better performance on the MS COCO dataset Lin et al. (2014), demonstrating the superior generalization ability of vision transformers on human pose estimation tasks even without multi-resolution features or cascade refinement. Besides, we investigate the scalability and flexibility of ViTPose regarding model size, input resolution, and token number for better performance. We also show that ViTpose can be easily pretrained using the unlabeled pose data without the need for the large-scale upstream ImageNet data Deng et al. (2009). ViTPose obtains the best 81.1 mAP on the MS COCO test-dev set.

2 Method

Simple vision transformer baselines. In this paper, we try to explore the potential of the plain vision transformer for human pose estimation without elaborate modifications. Thus, following Xiao et al. (2018), we directly feed the input images $I \in \mathcal{R}^{H \times W \times 3}$ into vision transformers to extract the features $F \in \mathcal{R}^{H/d \times W/d \times C}$ from the last transformer block, where d is the downsampling ratio of the vision transformer and C is the channel dimension. Typically, $d = 16$ for ViT variants. To align with the routine of human pose estimation methods, the feature maps are further upsampled to $1/4$ input size using several deconvolution layers. The channel dimension is downsampled to 256 to reduce computational costs for each deconvolution layer. Such a simple baseline with the ViT-Base backbone and 256×192 input resolution obtains 75.8 mAP on the MS COCO validation set, which outperforms or is comparable with the state-of-the-art (SOTA) results based on vision transformers, e.g., 75.6 mAP from HRFormer, and 75.8 mAP from TokenPose and TransPose.

Pretraining. Vision transformers are usually data-hungry and need a lot of training data for better performance. Many works have been proposed to address this problem. In this paper, we initialize the vision transformer backbones with MAE He et al. (2021) pretrained weights on ImageNet-1K Deng et al. (2009), which contains 1M image data. On the other hand, there are about 500K person instances in the human pose datasets such as MS COCO Lin et al. (2014) and AI Challenger Wu et al. (2017). We wonder whether ImageNet-1K pretraining is necessary for better human pose estimation. Surprisingly, using the unlabelled pose data only for MAE pretraining, the simple ViTPose baselines get similar performance compared with using backbones pretrained on ImageNet-1K.

Finer-resolution feature maps. To further improve the performance of simple vision transformer baselines, we explore the benefits of using finer-resolution feature maps for pose estimation. Specifically, assume the downsampling ratio d of ViT is 16, ViT first downsamples the input image to 1/16 of its input size by partitioning the image into non-overlapping windows, where the size of each window matches the downsampling ratio. Then, each window is mapped to a feature vector (i.e., token) with channel dimension C and processed by the transformer blocks to model global dependencies. To generate finer-resolution feature maps with minimal modifications to the plain visual transformer, we simply use overlapping windows with padding to partition the input image. For example, to increase the downsampling ratio from 16 to 8, we partition the image with a window size of 16 × 16 and an overlap of 8 pixels between adjacent windows. Although such an operation can help us generate finer-resolution feature maps (i.e., more tokens), using global transformer blocks to process the finer feature maps will bring extra computations and an extremely huge memory footprint. It is because the computation complexity of the naive self-attention in transformers is squared to the number of tokens. We resort to a full-window attention structure to reduce the memory load to address this issue. Nevertheless, window attention without global propagation heavily downgrades ViT's performance since it can not model the global dependencies. To enable the information propagation between different windows, a simple modification of ViT is adopted, i.e., the ViTAE transformer Xu et al. (2021), which employs parallel convolution branches beside the attention blocks to aid the information communication between adjacent windows.

3 Experiment

Experiment settings. We use the COCO Keypoint detection dataset Lin et al. (2014) for training and test. The training set contains 150k person instances with 118k images. The val and test-dev set of the COCO dataset contains 5k and 20k images for evaluation. For performance evaluation on the val set, we use the person detection results from Xiao et al. (2018) to align with previous methods. A stronger detector Cai et al. (2022) is used to detect person instances on the test-dev set, which achieves 68.5 mAP for the person class on the COCO val set.

3.1 Ablation Study

We adopt ViT-B pretrained by MAE on the ImageNet dataset Deng et al. (2009) as the baseline in the ablation study. The models are further trained on the training images from the COCO datasets for 210 epochs, and the performance is reported on the val set of the COCO dataset with person detection results provided by Xiao et al. (2018).

3.1.1 The influence of different resolutions

Table 1: The performance of ViT-B using different resolutions on COCO val set.

	224x224	256x192	256x256	384x288	384x384	576x432
AP	74.9	75.8	75.8	76.9	77.1	77.8
AR	80.4	81.1	81.1	81.9	82.0	82.6

To evaluate the flexibility and performance of ViTPose regarding different input resolutions, we resize the input images to different resolutions and use these images to train and test the simple ViTPose baseline. The results are summarized in Table 1. Generally, with larger input images, ViTPose can generate better performance for the human pose, e.g., with 576×432 input resolution, ViTPose obtains an absolute 2 mAP increase compared with using the 256×192 input resolution for training and test. Besides, using square inputs does not bring much benefit to the pose estimation task than using rectangular inputs, e.g., the performances are almost the same either using 256×256 or 256×192 as input resolution as well as using 384×384 or 384×288 as input resolution.

3.1.2 The influence of pretraining data

Table 2: The performance of ViT-B using different data for MAE pretraining on COCO val set.

Pretraining Dataset	Dataset size	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
ImageNet-1k	1M	75.8	90.7	83.2	81.1	94.6	87.7
COCO+AI Challenger	500k	75.8	90.8	83.0	81.0	94.6	87.4

To alleviate the data-hungry issue of vision transformers, we use MAE pretrained vision transformers to initialize the backbone networks for human pose estimation. However, the pretraining method requires extra data, i.e., ImageNet-1K, for pose estimation. We wonder whether it is necessary to use such a large-scale dataset for pretraining. To this end, we train a ViT-B model from scratch using MAE on the unlabelled person instances from COCO and AI Challenger human pose datasets for 1600 epochs following the same setting in He et al. (2021). Then, the pretrained model is used to initialize the backbone for pose estimation and finetuned with training images from COCO for 210 epochs. As shown in Table 2, models pretrained on human pose data obtain almost the same performance as the ImageNet-1K pretrained models. The result demonstrates that if there are enough images in the downstream tasks, there is no need to use extra training data from upstream tasks for pretraining, and the domain-specific data themselves can provide a good initialization.

3.1.3 The influence of finer-resolution feature maps

Table 3: The performance of ViT-B and ViTAE-B with full-window attention and 1/8 downsampling ratio on COCO val set.

	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
ViT-B w/ Full-Window	72.0	89.9	78.0	77.7	93.7	83.4
ViTAE-B w/ Full-Window	76.3	90.9	82.9	81.4	94.6	87.3

To evaluate the performance of ViTPose with full-window attention on finer-resolution feature maps, we change the downsampling ratio of pretrained models from 16 to 8 and use 256×192 images during training. As demonstrated in Table 3, the ViT-B model with full-window attention achieves much worse results on the pose estimation tasks, compared with 75.8 mAP of the naive global attention version in Table 1. The result demonstrates that global information propagation is important for pose estimation. With parallel convolution branches for information exchange between adjacent windows, the ViTAE-B model with full-window attention achieves much better performance, e.g., 76.3 mAP v.s. 72.0 mAP.

3.1.4 The influence of backbone model size

Table 4: The performance of ViTPose with ViT-B, ViT-L, and ViT-H backbones on COCO val set.

Backbone	#Parameter	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
ViT-B	86M	75.8	90.7	83.2	81.1	94.6	87.7
ViT-L	307M	78.3	91.4	85.2	83.5	95.3	89.5
ViT-H	632M	79.1	91.7	85.7	84.1	95.4	90.0

As demonstrated in other vision tasks like image classification or object detection Dosovitskiy et al. (2020); Li et al. (2022), ViT models can be easily scaled to larger sizes. We also evaluate the scalability of ViTPose by adopting ViT-B, ViT-L, and ViT-H as the backbone networks for human pose estimation. As shown in Table 4, the average precision and recall increase steadily as the model size becomes larger, demonstrating the excellent scalability of ViTPose regarding model size.

3.2 Comparison with SOTA methods

Table 5: Comparison between ViTPose and SOTA methods on the COCO test-dev set. "+" denotes ensemble models. "†", "‡", and "*" denote the champion of the 2018, 2019, and 2020 COCO Keypoint Challenge, respectively.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_{M}	AP_{L}	AR
Baseline^{+} Xiao et al. (2018)	ResNet-152	76.5	92.4	84.0	73.0	82.7	81.5
HRNet Sun et al. (2019)	HRNet-w48	77.0	92.7	84.5	73.4	83.1	82.0
MSPN$^{+\dagger}$ Li et al. (2019)	4xResNet-50	78.1	94.1	85.9	74.5	83.3	83.1
DARK Zhang et al. (2020)	HRNet-w48	77.4	92.6	84.6	73.6	83.7	82.3
RSN$^{+\ddagger}$ Cai et al. (2020)	4xRSN-50	79.2	94.4	87.1	76.1	83.8	84.1
CCM^{+} Zhang et al. (2021)	HRNet-w48	78.9	93.8	86.0	75.0	84.5	83.6
UDP++$^{+*}$ Huang et al. (2020)	HRNet-w48plus	80.8	94.9	88.1	77.4	85.7	85.3
TransPose-H-A6 Yang et al. (2021)	HRNet-w48	75.0	92.2	82.3	71.3	81.1	-
TokenPose-L/D24 Li et al. (2021b)	HRNet-w48	75.9	92.3	83.4	72.2	82.1	80.8
HRFormer Yuan et al. (2021)	HRFormer-B	76.2	92.7	83.8	72.5	82.3	81.2
ViTPose	ViTAE-G	80.9	94.8	88.1	77.5	85.9	85.4
ViTPose^{+}	**ViTAE-G**	**81.1**	**95.0**	**88.2**	**77.8**	**86.0**	**85.6**

We provide the comparison between ViTPose and SOTA methods on the COCO test-dev set. As demonstrated in Tabel 5, ViTPose with a single ViTAE-G backbone has obtained the best performance on the COCO test-dev set, outperforming all previous champions including UDP++, which uses a stronger detector with 68.6 mAP for the person class and ensembles 17 different models. The ensemble version of our four ViTPose models further obtains 81.1 mAP on the COCO test-dev set, setting the new SOTA for the human pose estimation task.

4 Conclusion

In this paper, we demonstrate that the plain vision transformer can generalize well on the human pose estimation task, even without specific designs. The proposed simple but effective vision transformer baseline, i.e., ViTPose, obtains the best 81.1 mAP on the COCO test-dev set, benefiting from bigger model size, larger input resolution, and more token numbers. We believe ViTPose can be also applied in other pose estimation tasks, e.g., animal pose estimation Yu et al. (2021). We hope this study can provide useful insights to the community and draw more attention on adapting the plain vision transformers for more vision tasks.

References

Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Mu Li, and Xiangyang Xue. Bigdetection: A large-scale benchmark for improved object detector pre-training. *arXiv preprint arXiv:2203.13249*, 2022.

Yuanhao Cai, Zhicheng Wang, Zhengxiong Luo, Binyi Yin, Angang Du, Haoqian Wang, Xinyu Zhou, Erjin Zhou, Xiangyu Zhang, and Jian Sun. Learning delicate local representations for multi-person pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1944–1953, June 2021a.

Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022.

Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, 2019.

Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660, 2014.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481, 2018.

Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in Neural Information Processing Systems*, 34, 2021.

Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. In *NeurIPS*, 2021.

Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7093–7102, 2020.

Jing Zhang and Dacheng Tao. Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things. *IEEE Internet of Things Journal*, 8(10):7789–7817, 2020.

Jing Zhang, Zhe Chen, and Dacheng Tao. Towards high performance human keypoint detection. *International Journal of Computer Vision*, 129(9):2639–2662, 2021.

Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *arXiv preprint arXiv:2202.10108*, 2022a.

Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vsa: Learning varied-size window attention in vision transformers. *arXiv preprint arXiv:2204.08446*, 2022b.