Convolution Neural Network and Transformer-Based Human Pose Estimation

Subjects: Computer Science, Artificial Intelligence

Contributor: Chengyu Wu , Xin Wei , Shaohua Li , Ao Zhan

Human pose estimation is a complex detection task in which the network needs to capture the rich information contained in the images.

human pose estimation multi-scale Transformer coordinate attention

1. Introduction

Human pose estimation is a crucial component in the field of computer vision, aiming to predict anatomical keypoints of the human body in 2D images. With the advancement of deep convolutional neural networks, the performance of pose estimation models has made significant progress. These models have gradually been applied to more complex scenarios, such as motion analysis [1][2][3] and human–computer interaction [4][5][6].

Currently, the mainstream models for pose estimation predominantly rely on convolution neural network (CNN) as encoders to extract texture features. Subsequently, the feature maps are decoded into higher-resolution sizes using methods, such as heatmap-based approaches or direct keypoint regression, which has become a widely adopted paradigm in most pose estimation models. The Hourglass model [7], for instance, stacks multiple Hourglass modules, where each module utilizes symmetric up-sampling and down-sampling processes combined with intermediate supervision to generate high-resolution feature maps. HRNet [8] employs parallel branches for different resolution feature mappings while consistently maintaining the highest-resolution branch. However, due to the nature of convolutional kernels, CNN exhibits local convolutional properties, restricting its ability to capture global dependencies within a limited receptive field. Although CNN excels at extracting texture features from images, it often lacks the capacity to learn spatial features effectively. As a consequence, the network fails to fully comprehend the information contained in the image. These limitations greatly constrain the potential of CNN-based models.

In recent years, Transformer [9] has achieved remarkable success in the field of natural language processing (NLP), continuously breaking records and topping various leaderboards. As a sequence-to-sequence model, Transformer exhibits strong modeling capabilities for dependencies between sequences. Furthermore, in the field of computer vision, Transformer excels at capturing the spatial features of images. The introduction of Vision Transformer (ViT) [10] marked the first application of Transformer to computer vision. The authors divided images into smaller patches, flattened them into sequences, and trained Transformer on these sequences. This simple yet effective approach quickly attracted the attention of many researchers. However, the high resolution of images

poses computational challenges for pure Transformer methods, leading to the emergence of CNN + Transformer networks. One of the most representative models in this category is TFPose [11]. The authors employ CNN as an encoder, flatten the extracted features along the channel dimension, and feed them into Transformer. Finally, they use regression methods to predict keypoints. Researchers believe that CNN + Transformer is a more optimal solution that leverages the strengths of both networks, striking a balance between speed and accuracy. However, mainstream CNN + Transformer models are still in their early stages, leaving room for exploration in terms of network integration and regression approaches. Consequently, the performance of both networks is not fully realized. Based on these considerations, this research proposes a novel network architecture called MSTPose, which aims to address the limitations of existing models.

2. Convolution Neural Network-Based Human Pose Estimation

In the field of human pose estimation, CNN-based methods have achieved tremendous success. Many early works aim to extract image features by using CNN as an encoder. DeepPose [12] firstly introduces CNN to address the problem of pose estimation, they propose a cascaded structure of deep neural networks. In SimpleBaseline [13], the authors utilize transpose convolution in the output part of the backbone network to generate higher-resolution feature maps for better pose estimation.

Due to pose estimation being different from simple detection tasks, capturing the global dependencies between features is crucial. Varun Ramakrishna et al. [14] propose a sequential prediction algorithm that simulates the mechanism of message passing to predict the confidence of each variable (part), iteratively improving the estimation at each stage. Tompson et al. [15] utilize the structural relationships between human keypoints and incorporate the idea of Markov random fields to optimize the prediction results. Wei et al. [16] introduce the CPM (convolutional pose machines) network with VGG [17] as the backbone, employing a jointly trained multi-stage, intermediate supervision architecture to learn the dependencies between keypoints. George Papandreou et al. [18] propose a box-free system based on fully convolutional networks, learning the offsets of keypoints through a greedy decoding process and grouping keypoints into human pose instances.

However, due to the local convolutional nature of CNN, its ability to capture global dependencies is limited. Another approach is to enlarge the receptive field of feature maps, and there are various ways to achieve this, such as multi-scale fusion [19][20][21][22] and high-resolution representation [23]. Yilun Chen et al. [21] present a cascaded pyramid model to obtain multi-scale features, ultimately performing pose estimation by up-sampling to high-resolution feature maps. Bowen Cheng et al. [23] propose HigherNRNet, which utilizes transpose convolutions to obtain higher-resolution feature maps to perceive small-scale objects.

As networks become increasingly complex, there is a need for better methods to more comprehensively capture image information. Compared to previous works that solely rely on CNN, the emergence of Transformer highlights new possibilities to pose estimation.

3. Transformer-Based Human Pose Estimation

Transformer is a feed-forward network based on the self-attention mechanism, which has achieved significant success in the field of NLP [24][25][26][27][28][29]. In recent years, with the introduction of Transformer into the visual domain, researchers have witnessed the rise of Transformer [10][30][31].

In the field of image segmentation, W. Wang et al. [32] propose a method called Attention-Guided Object Segmentation (AGOS) and Dynamic Visual Attention Prediction (DVAP) for unsupervised video object segmentation. T. Zhou et al. [33] introduce Matnet, which employs a two-stream encoder to transform surface features into mobile attention features at each stage of convolution. The bridge network is used for multi-level feature map fusion and acquisition, resulting in better segmentation results.

In the domain of object detection, N. Carion et al. [30] present the DETR model, which achieves higher detection accuracy by incorporating Transformer and employing a unique set prediction loss. To address the slow convergence speed and limited feature spatial resolution issues in [30], X. Zhu et al. [31] propose Deformable DETR, where the attention module focuses only on a small group of key sampling points around the reference, leading to improved performance.

In the field of human pose estimation, S. Yang et al. [34] introduce TransPose, using CNN as the encoder and incorporating Transformer for precise localization of human keypoints, capturing both short and long-range dependencies between keypoints. W. Mao et al. [11] propose TFPose, which builds upon [34] and employs direct regression of keypoints for pose estimation. K. Li et al. [35] develop the end-to-end PRTR model, which employs cascaded Transformer networks for direct regression of keypoints. B. Shan et al. [36] propose the MSRT network, which performs segmentation and superimposition of feature maps at different scales using the FAM module and utilizes Transformer for keypoints decoding.

References

1. Meng, Z.; Zhang, M.; Guo, C.; Fan, Q.; Zhang, H.; Gao, N.; Zhang, Z. Recent Progress in Sensing and Computing Techniques for Human Activity Recognition and Motion Analysis. Electronics 2020, 9, 1357.
2. Agostinelli, T.; Generosi, A.; Ceccacci, S.; Khamaisi, R.K.; Peruzzini, M.; Mengoni, M. Preliminary Validation of a Low-Cost Motion Analysis System Based on RGB Cameras to Support the Evaluation of Postural Risk Assessment. Appl. Sci. 2021, 11, 10645.
3. Maskeliūnas, R.; Damaševičius, R.; Blažauskas, T.; Canbulut, C.; Adomavičienė, A.; Griškevičius, J. BiomacVR: A Virtual Reality-Based System for Precise Human Posture and Motion Analysis in Rehabilitation Exercises Using Depth Sensors. Electronics 2023, 12, 339.

4. Liu, H.; Liu, T.; Zhang, Z.; Sangaiah, A.K.; Yang, B.; Li, Y. ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction. IEEE Trans. Ind. Inform. 2022, 18, 7107–7117.
5. Liu, H.; Li, D.; Wang, X.; Liu, L.; Zhang, Z.; Subramanian, S. Precise head pose estimation on HPD5A database for attention recognition based on convolutional neural network in human-computer interaction. Infrared Phys. Technol. 2021, 116, 103740.
6. Wang, K.; Zhao, R.; Ji, Q. Human computer interaction with head pose, eye gaze and body gestures. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG, Xi'an, China, 15–19 May 2018; p. 789.
7. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part VIII 14. pp. 483–499.
8. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4 December 2017; pp. 6000–6010.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
11. Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; Wang, Z. Tfpose: Direct human pose estimation with transformers. arXiv 2021, arXiv:2103.15320.
12. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
13. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 466–481.
14. Ramakrishna, V.; Munoz, D.; Hebert, M.; Andrew Bagnell, J.; Sheikh, Y. Pose machines: Articulated pose estimation via inference machines. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part II 13. pp. 33–47.
15. Tompson, J.J.; Jain, A.; LeCun, Y.; Bregler, C. Joint training of a convolutional network and a graphical model for human pose estimation. In Proceedings of the Advances in Neural Information

Processing Systems, Long Beach, CA, USA, 4 December 2017.

16. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4724–4732.
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
18. Papandreou, G.; Zhu, T.; Chen, L.C.; Gidaris, S.; Tompson, J.; Murphy, K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 269–286.
19. Pfister, T.; Charles, J.; Zisserman, A. Flowing convnets for human pose estimation in videos. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1913–1921.
20. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning feature pyramids for human pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1281–1290.
21. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7103–7112.
22. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1840.
23. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5386–5395.
24. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805.
25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. arXiv 2019, arXiv:1907.11692.
26. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 14 October 2022).

27. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv 2019, arXiv:1910.13461.
28. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 2020, 21, 5485–5551.
29. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X.V.; et al. Opt: Open pre-trained transformer language models. arXiv 2022, arXiv:2205.01068.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part I 16. pp. 213–229.
31. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. arXiv 2020, arXiv:2010.04159.
32. Wang, W.; Song, H.; Zhao, S.; Shen, J.; Zhao, S.; Hoi, S.C.; Ling, H. Learning unsupervised video object segmentation through visual attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3064–3074.
33. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. Matnet: Motion-attentive transition network for zero-shot video object segmentation. IEEE Trans. Image Process. 2020, 29, 8326–8338.
34. Yang, S.; Quan, Z.; Nie, M.; Yang, W. Transpose: Keypoint localization via transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11802–11812.
35. Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z. Pose recognition with cascade transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 19–25 June 2021; pp. 1944–1953.
36. Shan, B.; Shi, Q.; Yang, F. MSRT: Multi-scale representation transformer for regression-based human pose estimation. Pattern Anal. Appl. 2023, 26, 591–603.