Efficient Human Body Pose Estimation with Patch Selection

Kaleab A. Kinfu[1] and René Vidal[2]

[1]Mathematical Institute for Data Science, Johns Hopkins University, [2]Center for Innovation in Data Engineering and Science, University of Pennsylvania

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

- Human pose estimation has become crucial in computer vision applications like surveillance, autonomous driving, and augmented reality.
- While CNNs have been successful for this task, Vision Transformers (ViTs) have emerged as a powerful alternative.
- However, the quadratic computational complexity of ViTs motivates the need to reduce the number of tokens that need to be processed.

Related Work

- Recent work [1] shows that not all image patches contribute equally to the human pose estimation task because long-range dependencies relevant to predicting keypoints are primarily confined to the body part regions.
- Therefore, performing multi-head self-attention computations between every patch in the image becomes unnecessary.

Contributions

- We propose a transformer-based human pose estimation approach that uses a lightweight pose estimation network to select a few patches, which are processed by a ViT encoder and a CNN decoder that predicts the pose.
- Experiments on three datasets show that our method significantly reduces the computational overhead while maintaining a good level of accuracy.

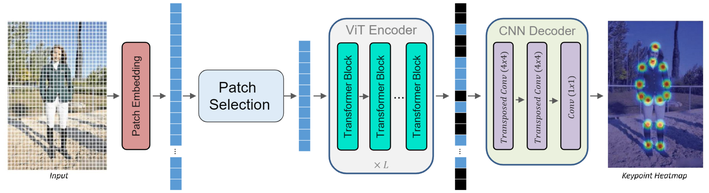


Figure 1: Overall architecture of the proposed ViT-based HPE method with patch selection -- The input is fed to a patch embedding layer that divides the image into patches of size 16×16. Patch selection is performed before they are processed by ViT to reduce the computation. Then, a simple CNN decoder is fed with the featuremap generated by ViT, which includes zero-filled non-body-part patches, to generate the heatmap prediction.

Efficient Pose Estimation via Patch Selection

- We propose two methods that leverage a lightweight pose estimator to guide the selection of relevant body part patches while discarding irrelevant ones, reducing computational complexity.
  - The first method (see Algorithm 1) utilizes a breadth-first neighboring search algorithm to select body joint patches and its neighboring patches.
  - The second method (see Algorithm 2) focuses on selecting patches formed by the skeleton of joint connections. By utilizing Bresenham's algorithm, the approach identifies patches where the lines formed by body joint pairs intersect.
- These methods avoid the need for re-training the vision transformer and effectively reduce computational complexity by focusing on the small relevant patches using the lighweight pose estimator's guidance.

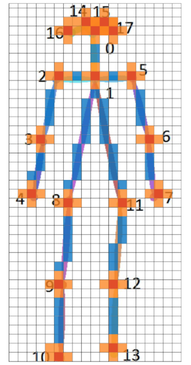
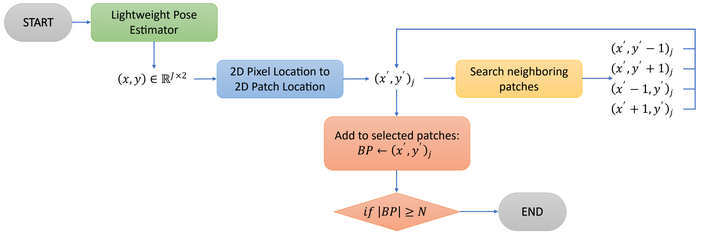
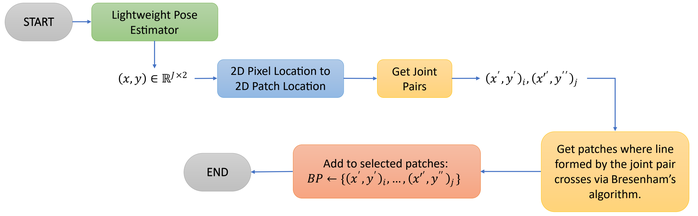


Figure 2. **Selected patches via the patch selection methods.** The red, orange, and blue patches correspond to the body keypoints, neighboring patches, and skeleton patches, respectively.



Algorithm 1: Neighboring patch selection method.



Algorithm 2: Skeleton-based patch selection method.

Results

Experiments on the MS COCO val set, the MPII test set, and the OCHuman test set demonstrate that our methods significantly improve speed and reduce computational complexity, with only a slight drop of 1-5% in accuracy.

Table 1. Performance of the proposed patch selection methods ($n = 7$) on three benchmarks, namely MS COCO, MPII, and OCHuman.

Model	Patch Selection	Input Resolution	Params	FLOPs	COCO mAP	MPII PCKh	OCHuman mAP
Lite-HRNet [21]	None	256×192	1M	0.2G	64.8	86.1	51.9
SimpleBaseline [17]	None	256×192	69M	15.7G	72.0	89.0	58.2
HRNet-W48 [16]	None	256×192	64M	14.6G	75.1	90.1	60.4
HRFormer-B [22]	None	256×192	43M	12.2G	75.6	-	49.7
ViTransPose-B	None	256×192	90M	17.9G	76.9	92.2	88.2
ViTransPose-B	Neighbors	256×192	90M	11.1G	73.4	91.5	84.7
ViTransPose-B	Skeleton	256×192	90M	13.3G	74.3	91.9	85.3
ViTransPose-L	None	256×192	309M	59.8G	78.7	92.8	91.5
ViTransPose-L	Neighbors	256×192	309M	35.6G	75.7	92.1	87.2
ViTransPose-L	Skeleton	256×192	309M	38.3G	76.3	92.4	89.8

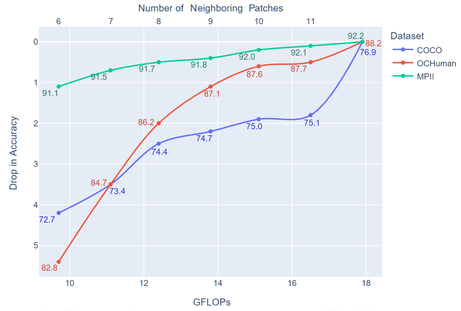


Figure 3. Trade-off between accuracy and GFLOPs of the neighboring patch selection method for the three benchmarks.

Limitations and extensions

- The reliance on performance of off-the-shelf pose estimators is a limitation.
- An extension work would involve developing an automatic patch selection method that overcomes the aforementioned limitation.

References

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