.

TĂNG CƯỜNG KHẢ NĂNG HỌC ĐẶC TRƯNG CỦA MÔ HÌNH HIGHERHRNET VỚI SE-BLOCK TRONG BÀI TOÁN ƯỚC LƯỢNG TƯ THẾ NGƯỜI

Phù Khắc Anh[1], Hoàng Văn Dũng[2*], Lê Văn Tường Lân[3]

[1]Khoa Công nghệ thông tin, Trường Đại học Khoa học, Đại học Huế

[2]Khoa Công nghệ thông tin, Trường Đại học Sư phạm kỹ thuật, Tp Hồ Chí Minh

[3]Đại học Huế

[*]Email: dunghv@hcmute.edu.vn

Ngày nhận bài: 22/11/2024; ngày hoàn thành phản biện: 16/12/2024; ngày duyệt đăng: 20/12/2024

TÓM TẮT

Rút trích dữ liệu khung xương hay nhận diện tư thế người là một tác vụ quan trọng trong các bài toán về nhận dạng hành động người hay theo dõi chuyển động. Nhiều nghiên cứu đã được thực hiện để đề xuất ra các mô hình khác nhau để có thể trích xuất dữ liệu khung xương từ dữ liệu hình ảnh, video. Trong nghiên cứu này, chúng tôi đề xuất một giải pháp nhằm nâng cao hiệu quả của mô hình. Nghiên cứu đề xuất tích hợp SE-Block vào mô hình HigherHRNet nhằm tăng khả năng học đặc trưng giữa các kênh trong bài toán trích xuất dữ liệu khung xương. Kết quả thực nghiệm trên tập COCO cho thấy mô hình cải tiến có sự ổn định tốt hơn trong quá trình huấn luyện và cải thiện nhẹ về chỉ số AR trong các tình huống tư thế phức tạp, qua đó mở ra nhiều hướng nghiên cứu mới trong việc nâng cao hiệu quả rút trích dữ liệu khung xương cũng như nhận dạng hành động hay theo dõi chuyển động của đối tượng.

Từ khóa: thị giác máy tính, học sâu, dữ liệu khung xương.

1. MỞ ĐẦU

Nhận dạng hành động của con người là một trong những bài toán được nhiều nhà nghiên cứu quan tâm trong lĩnh vực thị giác máy tính với nhiều ứng dụng thực tế như trong giám sát an ninh, giao tiếp giữa người với máy, y tế, thể thao, giải trí [1,2]. Ước lượng tư thế người – hay còn gọi là trích xuất dữ liệu khung xương – là quá trình xác định chính xác vị trí của các điểm khớp trên cơ thể người từ ảnh tĩnh hoặc chuỗi video. Các phương pháp ước lượng tư thế hiện nay chủ yếu được phát triển theo hai hướng tiếp cận chính: top-down và bottom-up [3]. Trong hướng tiếp cận top-down, mô hình đầu tiên sử dụng một thuật toán phát hiện đối tượng – chẳng hạn như Faster R-

CNN [4] hoặc YOLO [5] – để xác định các vùng chứa người trong ảnh. Sau đó, một mạng con sẽ được áp dụng cho từng vùng này để dự đoán vị trí các khớp tương ứng. Các mô hình nổi bật theo hướng tiếp cận này bao gồm Hourglass [6], HRNet [7] và ViTPose [8].

Ngược lại, hướng tiếp cận bottom-up thực hiện dự đoán tất cả các điểm khớp có thể có trong ảnh mà không cần tách riêng từng người. Sau khi dự đoán, các thuật toán phân nhóm sẽ được sử dụng để gán các khớp về từng đối tượng tương ứng. Một số mô hình tiêu biểu thuộc hướng tiếp cận này bao gồm OpenPose [9], OmniPose [10], MoveNet [11], v.v…Trong số các kiến trúc hiện đại, HigherHRNet [12] được xem là một bản cải tiến đáng chú ý của HRNet, vốn nổi bật với chiến lược duy trì độ phân giải cao xuyên suốt quá trình trích xuất đặc trưng. HigherHRNet bổ sung cơ chế phát sinh heatmap ở các tầng trung gian có độ phân giải thấp hơn, từ đó nâng cao độ chính xác trong định vị khớp xương, đặc biệt trong các trường hợp tư thế phức tạp.

Tuy nhiên, hạn chế của HigherHRNet – cũng như nhiều mô hình tích chập truyền thống – là chưa có cơ chế điều chỉnh động mức độ quan trọng giữa các kênh đặc trưng. Đây chính là lý do mà ở nghiên cứu này, chúng tôi hướng đến khắc phục thông qua việc tích hợp SE-Block [13] vào cấu trúc mô hình. Chúng tôi chọn SE-Block thay vì các cơ chế attention khác vì SE-Block có cấu trúc đơn giản, dễ tích hợp và đã được chứng minh hiệu quả trên nhiều mô hình backbone tiêu chuẩn như ResNet và Inception [13]. Các nghiên cứu [14, 15] đã tích hợp SE-Block vào HRNet cho các bài toán như phân đoạn ngữ nghĩa hoặc phân loại cảnh viễn thám, và cho thấy kết quả cải thiện rõ rệt mà không làm tăng đáng kể chi phí tính toán. Do HigherHRNet là kiến trúc mở rộng từ HRNet, việc sử dụng một cơ chế attention đã tương thích tốt với HRNet là lựa chọn hợp lý và ổn định hơn so với các mô-đun khác. Đóng góp chính của nghiên cứu này được tóm lược như sau:

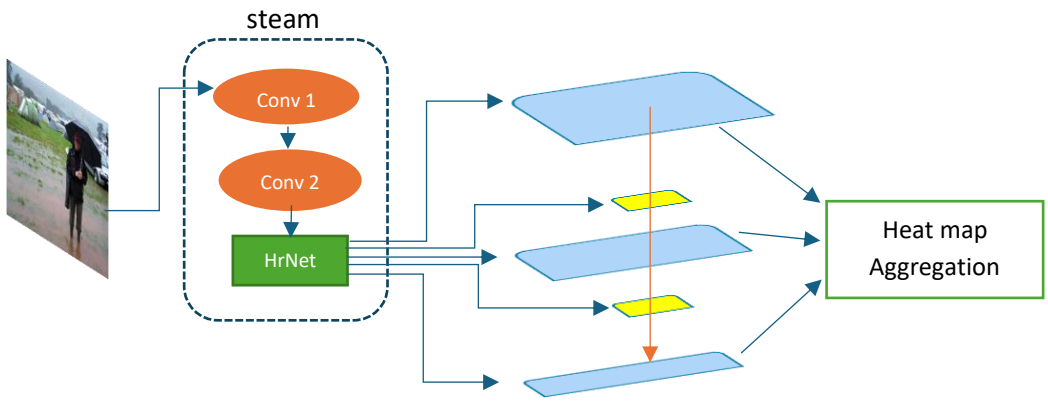
- Thứ nhất, chúng tôi tích hợp SE-Block vào giữa hai lớp tích chập đầu tiên trong mô-đun trích xuất đặc trưng ban đầu của HigherHRNet, giúp mô hình học được mức độ quan trọng tương đối giữa các kênh đặc trưng, từ đó làm nổi bật thông tin có giá trị và giảm ảnh hưởng từ nhiễu.
- Thứ hai, mô hình cải tiến được huấn luyện và đánh giá trên tập dữ liệu COCO [16] – một tập dữ liệu phổ biến cho bài toán ước lượng tư thế người – và được so sánh trực tiếp với phiên bản HigherHRNet gốc trong cùng điều kiện huấn luyện. Kết quả thực nghiệm cho thấy việc tích hợp SE-Block mang lại cải thiện nhất định về độ chính xác, đặc biệt trong các trường hợp khó như khớp bị che khuất hoặc chồng lấp.

2. NGHIÊN CỨU LIÊN QUAN

HigherHRNet là một kiến trúc mở rộng từ HRNet (High-Resolution Network), được thiết kế nhằm nâng cao hiệu quả trong các bài toán ước lượng tư thế người, đặc

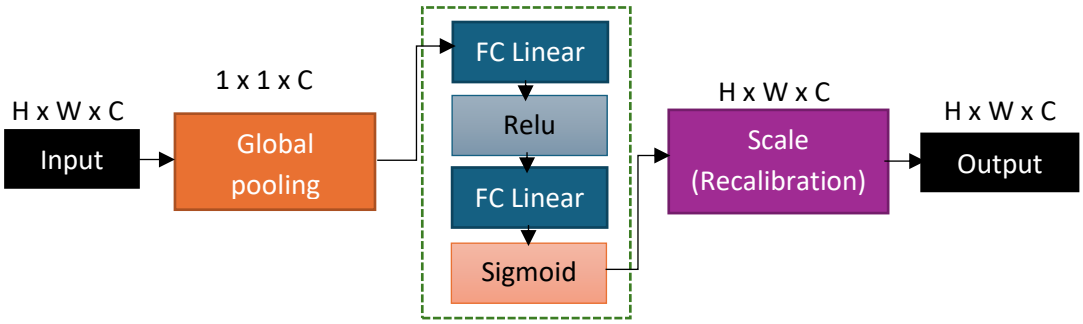
biệt là trong việc phát hiện các điểm khớp với độ chính xác cao. Trong khi hầu hết các kiến trúc truyền thống thực hiện quá trình giảm độ phân giải rồi sau đó khôi phục lại bằng các tầng upsampling, thì HRNet – và mở rộng hơn là HigherHRNet – duy trì độ phân giải cao xuyên suốt toàn bộ kiến trúc, đồng thời kết hợp thông tin từ nhiều tầng độ phân giải khác nhau một cách tuần tự và hiệu quả.

Điểm đặc biệt của HigherHRNet nằm ở khả năng xây dựng heatmap ở các cấp độ phân giải khác nhau, tức là mô hình không chỉ dự đoán các điểm khớp ở cấp độ đầu ra cuối cùng mà còn sinh ra các heatmap phụ tại các tầng trung gian có độ phân giải thấp hơn. Cấu trúc của HigherHRNet bao gồm các mô-đun High-Resolution Modules, trong đó mỗi nhánh xử lý một độ phân giải khác nhau, và các mô-đun này được kết nối lặp lại tuần hoàn để làm giàu đặc trưng không gian.



Hình 1. Mô hình HigherHRNet

Với các đặc điểm thiết kế như vậy, HigherHRNet đã chứng minh hiệu quả vượt trội trên nhiều tập dữ liệu chuẩn như COCO, MPII [17] và CrowdPose [18]. Tuy nhiên, như nhiều kiến trúc CNN truyền thống khác, HigherHRNet vẫn xử lý các kênh đặc trưng một cách đồng đều mà không xét đến mức độ quan trọng khác nhau giữa chúng. Điều này đặt ra nhu cầu tích hợp các cơ chế chú ý để tăng cường khả năng học biểu diễn, từ đó mở ra hướng cải tiến hiệu quả mà nghiên cứu này đề xuất thông qua việc kết hợp SE-Block.



Hình 2. Kiến trúc khối SE-Block

Bộ dữ liệu COCO (Common Objects in Context) là một trong những tập dữ liệu phổ biến và có ảnh hưởng lớn trong lĩnh vực thị giác máy tính, được sử dụng rộng rãi cho các bài toán như phát hiện đối tượng, phân đoạn ảnh, và đặc biệt là ước lượng tư thế người. Trong nghiên cứu này, chúng tôi sử dụng tập huấn luyện và tập kiểm tra của COCO để huấn luyện và đánh giá mô hình đề xuất.

SE-Block là một khối module có kiến trúc đơn giản nhưng hiệu quả, được thiết kế để giúp các mô hình học sâu chú ý tốt hơn vào những kênh đặc trưng quan trọng. Ý tưởng chính của SE-Block là thay vì xử lý mọi kênh một cách giống nhau, mô hình nên ưu tiên những kênh chứa thông tin hữu ích và giảm bớt ảnh hưởng của các kênh không cần thiết. SE-Block có thể tích hợp rất linh hoạt vào nhiều kiến trúc CNN hiện có, và thường mang lại kết quả cải thiện rõ rệt mà không làm mô hình nặng thêm nhiều.

Một nghiên cứu liên quan là CE-HigherHRNet [19], trong đó CBAM [20] được tích hợp sau các tầng hợp nhất đa tỉ lệ nhằm tăng cường thông tin ngữ nghĩa. Cách triển khai này giúp cải thiện độ chính xác, nhưng cũng làm tăng độ phức tạp do CBAM kết hợp attention theo cả kênh và không gian, trong đó spatial attention sử dụng tích chập 7×7 trên toàn bộ bản đồ đặc trưng. Trong nghiên cứu này, chúng tôi đề xuất tích hợp attention sớm hơn – tại giai đoạn tiền xử lý đặc trưng – giữa hai lớp convolution đầu vào. Nếu áp dụng CBAM tại vị trí này, spatial attention sẽ xử lý các bản đồ đặc trưng có độ phân giải lớn, gây tăng chi phí tính toán. Ngược lại, SE-Block, với cơ chế chú ý đơn giản theo chiều kênh, là lựa chọn phù hợp hơn để tăng cường biểu diễn ban đầu mà không làm phức tạp kiến trúc.

3. GIẢI PHÁP TÍCH HỢP SE-BLOCK VÀO HIGHERHRNET

Chúng tôi chọn SE-Block để tích hợp vào HigherHRNet vì đây là một giải pháp đơn giản, dễ áp dụng và đã được kiểm chứng là có hiệu quả trong nhiều mô hình học sâu. Cơ chế hoạt động của SE-Block gồm ba bước chính: nén không gian (squeeze), kích hoạt kênh (excitation) và tái hiệu chỉnh đặc trưng (recalibration). Giả sử đầu vào tại một thời điểm là một tensor $X \in R^{C \times H \times W}$, trong đó: C là số lượng kênh đặc trưng (số feature map) – ví dụ như 64 hoặc 128 kênh đầu ra từ một lớp convolution, H là chiều cao của feature map (height), W là chiều rộng của feature map (width).

Nói cách khác, một tensor 3 chiều $C \times H \times W$ tương ứng với C feature map, với mỗi feature map có kích thước là $H \times W$. Quá trình Squeeze – tổng hợp thông tin theo không gian được thực hiện bằng cách tính trung bình toàn cục như sau:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \text{ với c= 1,2, …, C} \quad (1)$$

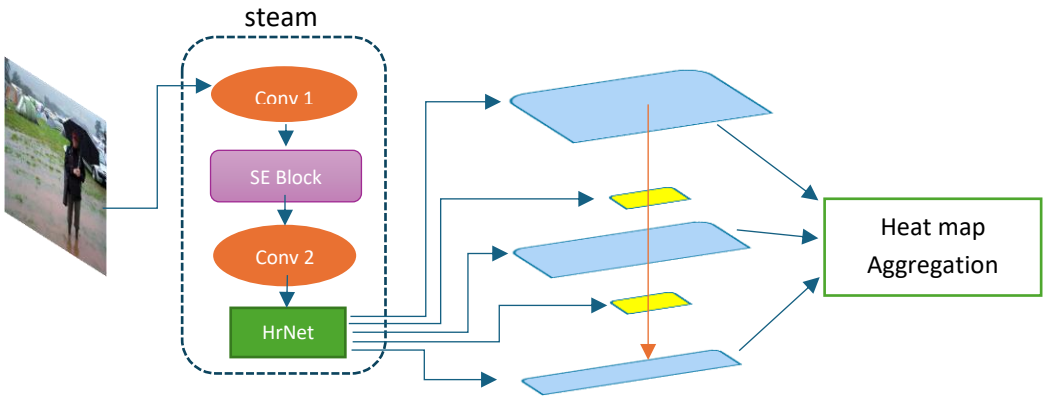
Vector z được đưa qua hai lớp fully connected với hàm ReLu và sigmoid để học trọng số $s \in R^C$ trong qúa trình Excitation:

$$s = \sigma(W_2 . \delta(W_1 . z)) \quad (2)$$

Với σ là hàm Relu, δ là hàm sigmoid, $W_1 \in R^{\frac{C}{r} \times C}, W_2 \in R^{C \times \frac{C}{r}}$, r là hệ số giảm chiều. Ở bước Recalibration, trọng số s được nhân với đầu vào X để tạo ra một đầu ra $\tilde{X}$ mới:

$$\tilde{X}_C = s_c . X_C \text{ với c = 1,2, …, C} \quad (3)$$

Trong quá trình thực nghiệm, chúng tôi huấn luyện mô hình HigherHRNet và phiên bản đã được tích hợp thêm SE-Block trên tập dữ liệu COCO. Máy dùng để huấn luyện có cấu hình gồm CPU Intel i5-13600K, RAM 64GB và GPU RTX 3090. Do giới hạn về phần cứng, chúng tôi không huấn luyện mô hình từ đầu mà sử dụng lại trọng số pretrained từ tác giả của mô hình HigherHRNet đã huấn luyện sẵn trên tập COCO với số epoch là 40, thời gian huấn luyện của cả 2 trường hợp là 36 giờ. Trong khi kiến trúc gốc của mô hình HigherHRNet được huấn luyện với tổng số epoch là 200. Chúng tôi tin rằng nếu trong điều kiện có thể tang số epoch huấn luyện lên đầy đủ hơn, hiệu quả của SE-Block mang lại sẽ thể hiện rõ hơn.



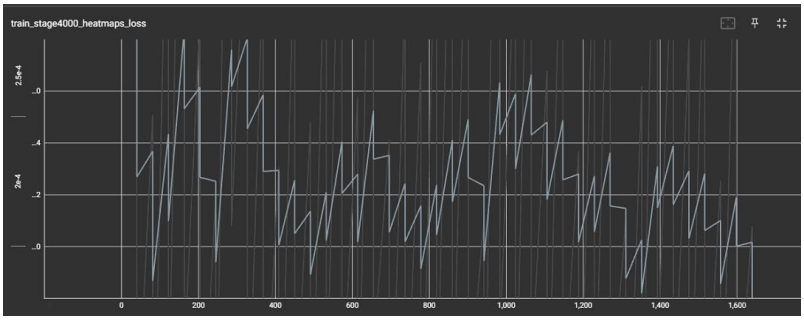
Hình 3. Sơ đồ tổng thể mô hình đề xuất với SE-Block tích hợp vào HigherHRNet

Bảng 1. Danh sách các tham số chính sử dụng trong quá trình huấn luyện

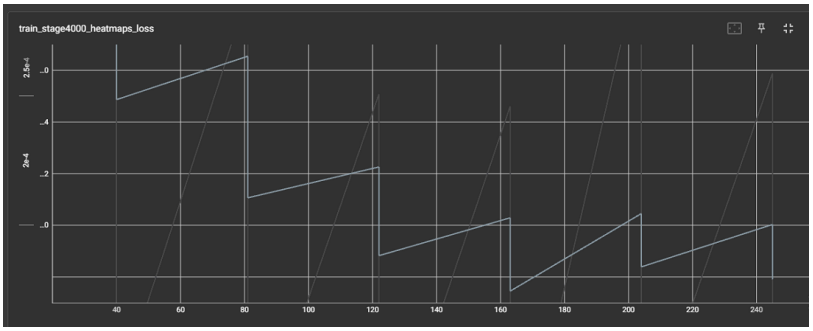
STT	Tham số	Giá trị	Ghi chú
1	Batch size	32	Kích thước mini-batch huấn luyện
2	Shuffle	True	Xáo trộn dữ liệu trong mỗi epoch
3	Optimizer	Adam	Bộ tối ưu hóa

4	Learning rate	0.001	Tốc độ học ban đầu
5	Learning rate step	[20, 30]	Các mốc epoch giảm learning rate
6	Learning rate factor	0.1	Hệ số giảm learning rate nếu không cải thiện
7	Số epoch huấn luyện	40	Tổng số vòng lặp huấn luyện
8	Use target weight	True	Sử dụng trọng số riêng cho từng điểm khớp trong hàm mất mát (loss)

Kết quả huấn luyện ở Hình 4 cho thấy mô hình tích hợp với SE-Block có độ ổn định tốt hơn đáng kể so với mô hình gốc. Cụ thể là tại train_stage4000 (stage cuối cùng của quá trình huấn luyện), heatmaps loss của mô hình gốc dao động mạnh mặc dù vẫn có xu hướng giảm, trong khi heatmaps loss của mô hình kết hợp với SE-Block tích hợp giảm đều và ít nhiễu hơn. Điều này cho thấy SE-Block đã giúp mô hình tập trung hiệu quả hơn vào các kênh đặc trưng quan trọng, qua đó góp phần cải thiện quá trình hội tụ của quá trình huấn luyện.



a) Heatmap loss của mô hình HigherHRNet



b) Heatmap loss của mô hình HigherHRNet có tích hợp SE – Block

Hình 4. Heatmaps loss của quá trình huấn luyện a) Mô hình HigherHRNet và b) Mô hình HigherHRNet có tích hợp SE – Block (smooth = 0.8)

Bảng 2. Kết quả valid mô hình HigherHRNet gốc và mô hình HigherHRNet có tích hợp SE –Block

Chỉ số	HigherHRNet	HigherHRNet + SE
AP	64.5%	64.3%
AP@0.5	85.0%	85.6%
AP@0.75	70.0%	69.9%
AP (M)	59.5%	59.4%
AP (L)	72.1%	**72.2%**
AR	69.5%	**69.7%**
AR@0.5	87.7%	**88.1%**
AR@0.75	74.0%	**74.1%**
AR (M)	63.4%	**63.5%**
AR (L)	78.3%	**78.5%**

a) b)

Hình 5. Kết quả valid của mô hình HigherHRNet gốc (Hình a) và mô hình HigherHRNet tích hợp với SE – Block – (Hình b)

Kết quả đánh giá cho thấy mô hình có SE-Block có sự cải thiện nhẹ về độ chính xác khi so với mô hình gốc, cụ thể ở các chỉ số AR như AR@0.5 và AR (L). Tuy chưa tạo

ra sự khác biệt rõ rệt ở giai đoạn này, nhưng quá trình huấn luyện ổn định hơn và đường loss mượt hơn cho thấy SE-Block đang giúp mô hình học tốt hơn về mặt lâu dài. Bên cạnh đó, thời gian huấn luyện gần như tương đương cũng như kiến trúc đơn giản của mình khiến SE-Block hoàn toàn xứng đáng được xem xét là một giải pháp phù hợp để cải tiến khi tích hợp vào các mô hình rút trích dữ liệu khung xương nói chung và mô hình HigherHRNet nói riêng.

Mặc dù SE-Block giúp ổn định quá trình hội tụ, một số chỉ số như AP@0.75, AP(M) giảm nhẹ. Điều này xuất phát hai nguyên nhân chính. Thứ nhất, SE-Block ở vị trí tích hợp trong nghiên cứu này cân bằng lại trọng số, ưu tiên các đặc trưng tổng quát nên đôi khi làm mờ những đặc trưng chi tiết cần cho các ngưỡng đánh giá ngưỡng khắt khe như AP@075. Thứ hai, với số epoch hiện tại còn hạn chế, mô hình chưa có đủ thời gian để học sâu các đặc trưng nhỏ này; khi tăng thêm epoch, AP ở các ngưỡng cao nhiều khả năng sẽ được cải thiện.

Vì kiến trúc HigherHRNet gốc thường được huấn luyện trong khoảng 200 epoch, chúng tôi tin rằng nếu tiếp tục huấn luyện với số epoch lớn hơn, hiệu quả mà SE-Block mang lại sẽ được thể hiện rõ hơn, đặc biệt trong các tình huống phức tạp như khớp bị che khuất hoặc ảnh có nhiều người. Hình 5 cho thấy mô hình HigherHRNet khi tích hợp SE – Block có thể phát hiện tốt hơn các khớp xương trong các trường hợp các bộ phận cơ thể người bị che khuất.

4. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã đề xuất cải tiến mô hình HigherHRNet bằng cách tích hợp thêm khối chú ý SE-Block nhằm tăng cường khả năng học các mối quan hệ không gian trong bài toán trích xuất dữ liệu khung xương người. Không thay đổi kiến trúc tổng thể, phương pháp đề xuất tập trung vào việc tăng hiệu quả xử lý đặc trưng theo kênh, từ đó cải thiện độ ổn định trong quá trình huấn luyện và chất lượng đầu ra.

Kết quả thực nghiệm trên tập dữ liệu COCO cho thấy mô hình có tích hợp SE-Block tăng nhẹ hiệu suất so với mô hình gốc tại mốc 40 epoch, với độ hội tụ ổn định hơn. Đặc biệt, trong các tình huống phức tạp như che khuất, mô hình có tích hợp SE-Block cho thấy khả năng nhận diện chính xác hơn. Tuy nhiên hạn chế trong nghiên cứu này là với số epoch huấn luyện còn hạn chế, SE-Block chưa học được các đặc trưng chi tiết, qua đó kết quả cải thiện của mô hình còn khiêm tốn. Chúng tôi tin rằng nếu tiếp tục huấn luyện với số epoch lớn hơn, hiệu quả của SE-Block sẽ càng được thể hiện rõ.

Trong quá trình thực nghiệm, chúng tôi cũng đã tiến hành thử tích hợp khối SE-Block vào tất cả các nhánh trong mô hình HigherHRNet. Tuy nhiên thời gian huấn luyện lại tăng lên đáng kể trong khi độ chính xác lại giảm. Vì vậy, trong tương lai, hướng nghiên cứu tiếp theo sẽ mở rộng việc áp dụng SE-Block có chọn lọc vào các nhánh phân

giải khác nhau trong HigherHRNet, cũng như kết hợp với các cơ chế chú ý nâng cao khác như CBAM để tiếp tục cải thiện chất lượng trích xuất khung xương trong các môi trường thực tế phức tạp. Bên cạnh đó việc tiến hành thực nghiệm trên các dataset khác như MPII, CrowdPose cũng như trên các mô hình có kiến trúc nhiều tầng khác cũng là một hướng nghiên cứu mà nhóm tác giả hướng đến.

TÀI LIỆU THAM KHẢO

[1]. M. H. Siddiqi et al., "A Unified Approach for Patient Activity Recognition in Healthcare Using Depth Camera," IEEE Access, vol. 9, pp. 92300–92317, 2021, doi: 10.1109/ACCESS.2021.3092403.

[2]. K. Kim, A. Jalal, and M. Mahmood, "Vision-Based Human Activity Recognition System Using Depth Silhouettes: A Smart Home System for Monitoring the Residents," Journal of Electrical Engineering & Technology, vol. 14, pp. 2567–2573, 2019.

[3]. P. K. Anh, H. V. Dũng, and L. V. T. Lân, "Khảo sát một số mô hình rút trích dữ liệu khung xương trong bài toán nhận diện hành động người," Tạp chí Khoa học và Công nghệ, Trường ĐH Khoa học – ĐH Huế, vol. 23, no. 1B, pp. 27–38, 2024.

[4]. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[5]. J. Terven and D. Cordova-Esparza, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," arXiv preprint, arXiv:2304.00501, 2023.

[6]. A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in Proc. ECCV, vol. 9912, pp. 483–499, 2016, doi: 10.1007/978-3-319-46484-8_29.

[7]. K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703, 2019. doi: 10.1109/CVPR.2019.00584.

[8]. Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation," in Proc. 36th Int. Conf. Neural Information Processing Systems (NeurIPS), Article no. 2795, pp. 38571–38584, 2022.

[9]. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[10]. B. Artacho and A. Savakis, "OmniPose: A Multi-Scale Framework for Multi-Person Pose Estimation," arXiv preprint, arXiv:2103.10180, 2021.

[11]. R. Votel and N. Li, "Next-Generation Pose Detection with MoveNet and TensorFlow.js," TensorFlow Blog, May 17, 2021. [Online]. Available: https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html

[12]. B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5386–5395, 2020. doi: 10.1109/CVPR42600.2020.00543.

[13]. J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

[14] L. Li, T. Tian, H. Li and L. Wang, "SE-HRNet: A Deep High-Resolution Network with Attention for Remote Sensing Scene Classification," IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 2020, pp. 533-536, doi: 10.1109/IGARSS39084.2020.9324633.

[15] Kim, Jin-Seong, Sung-Wook Park, Jun-Yeong Kim, Jun Park, Jun-Ho Huh, Se-Hoon Jung, and Chun-Bo Sim. 2023. "E-HRNet: Enhanced Semantic Segmentation Using Squeeze and Excitation" Electronics 12, no. 17: 3619. https://doi.org/10.3390/electronics12173619.

[16]. T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," arXiv preprint, arXiv:1405.0312, 2014.

[17]. M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in Proc. CVPR, 2014.

[18]. J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark," arXiv preprint, arXiv:1812.00324, 2019.

[19] M. Y. Li and J. Zhao, "CE-HigherHRNet: Enhancing Channel Information for Small Persons Bottom-Up Human Pose Estimation," IAENG International Journal of Computer Science, vol. 49, no. 1, pp. 27–34, Mar. 2022. [Online]. Available: https://www.iaeng.org/IJCS/issues_v49/issue_1/IJCS_49_1_27.pdf

[20] Woo, S., Park, J., Lee, JY., Kweon, I.S. (2018). CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(), vol 11211. Springer, Cham. https://doi.org/10.1007/978-3-030-01234-2_1

.

IMPROVING HIGHERHRNET'S FEATURE LEARNING WITH SE-BLOCK FOR HUMAN POSE ESTIMATION

Phu Khac Anh[1], Hoang Van Dung [2*], Le Van Tuong Lan[3]

[1] Faculty of Information Technology, University of Sciences, Hue University

[2] Faculty of Information Technology, HCMC University of Technology and Education

[3] Hue University

[*]Email: dunghv@hcmute.edu.vn

ABSTRACT

Skeleton extraction, or human pose estimation, is important role in tasks like action recognition and motion tracking. Over the years, many approaches have been developed to extract skeletal key points from images and video frames. This paper proposes a lightweight enhancement to HigherHRNet by integrating a Squeeze-and-Excitation (SE) block to adaptively recalibrate channel-wise features for better human pose estimation. Experiments on the COCO dataset show improved Average Recall (AR), especially in complex poses with occlusion and smoother loss convergence. This demonstrates the potential of SE-Block as a practical improvement for skeleton-based models, and it opens up possibilities for future work in pose estimation and related applications.

Keywords: computer vision, deep learning, skeleton data.

Phù Khắc Anh sinh ngày 01/03/1990 tại Long An. Ông tốt nghiệp cử nhân chuyên ngành Công nghệ thông tin tại trường Đại học Khoa Học Tự Nhiên, ĐH Quốc Gia Thành phố Hồ Chí Minh năm 2012, tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính năm 2016 tại trường Đại học Công Nghệ Thông Tin, ĐH Quốc Gia Thành phố Hồ Chí Minh năm 2016. Ông công tác tại trường Cao Đẳng Kỹ Thuật Cao Thắng Thành phố Hồ Chí Minh. Năm 2023 ông trở thành nghiên cứu sinh ngành Khoa học máy tính Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, thị giác máy tính

Hoàng Văn Dũng nhận bằng tiến sĩ ngành Kỹ thuật điện tử và Hệ thống thông tin tại Trường Đại học Ulsan, Hàn Quốc, năm 2015. Hiện nay, ông công tác tại Khoa Công nghệ thông tin, Trường ĐH Sư phạm Kỹ thuật TP. HCM.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, Thị giác máy tính

Lê Văn Tường Lân sinh ngày 10/11/1974 tại Thừa Thiên Huế. Năm 1996, ông tốt nghiệp Đại học ngành Toán - Tin tại Trường Đại học Khoa học, Đại học Huế. Ông nhận bằng thạc sỹ Công nghệ thông tin tại Trường Đại học Bách Khoa Hà Nội năm 2002 và nhận học vị Tiến sĩ ngành Khoa học máy tính tại Trường Đại học Khoa học, Đại học Huế năm 2018. Hiện nay, ông công tác tại Ban Đào tạo và Công tác sinh viên, Đại học Huế.

Lĩnh vực nghiên cứu: Cơ sở dữ liệu, Công nghệ phần mềm, Khai phá dữ liệu.