NGHIÊN CỨU MÔ PHỎNG DÁNG NGƯỜI TRÊN KHÔNG GIAN BA CHIỀU TỪ HÌNH ẢNH HAI CHIỀU SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU

RESEARCH IN RECREATING 3D HUMAN POSE FROM 2D IMAGES BY USING DEEP LEARNING

Phạm Lê Minh Hoàng*, Lê Thị Kim Oanh

Trường Đại học Bách khoa - Đại học Đà Nẵng[1]

*Tác giả liên hệ: plmhoang@dut.udn.vn

(Nhận bài: 15/02/2022; Chấp nhận đăng: 27/4/2022)

Tóm tắt - Nghiên cứu mô phỏng dáng người trong không gian ba chiều từ đơn ảnh đã có tiến triển đáng kể trong thời gian gần đây, nhờ tính toán bằng các mô hình có kiến trúc mạng tối ưu, kết hợp với các bộ dữ liệu quy mô lớn. Tuy nhiên, khi áp dụng vào điều kiện môi trường khác nhau trong thực tế, các phương pháp hiện có vẫn chưa đạt được độ chính xác so với kỳ vọng. Bài báo này đề xuất một giải pháp mới gồm hai mô hình kết hợp nhằm tăng độ chính xác dựa trên phương thức học sâu. Mô hình thứ nhất gọi là Squeeze-and-Excitation Network, được dùng để dựng lại dáng người hai chiều từ một ảnh đầu vào; Sau đó, sử dụng kết hợp giữa các lớp kết nối đầy đủ và mạng chập đồ thị để dựng thành dáng người ba chiều từ thông tin đầu ra của mô hình trước. Hiệu quả của phương pháp được chứng minh bằng cách so sánh với bộ dữ liệu chuẩn, và cho thấy độ chính xác được cải thiện đáng kể so với các phương pháp đã có trước.

Từ khóa - Mô phỏng dáng người; mô phỏng dáng người trong không gian ba chiều; đơn ảnh; mạng chập; học sâu.

Abstract - Recent studies have shown remarkable advances in 3D human pose estimation from monocular images, with the help of large-scale in-door 3D datasets and sophisticated network architectures. However, the expected generalizability to different environments remains an elusive goal to apply in the real-life tasks. In this work, we present a solution for single-view 3D human skeleton estimation based on deep learning method. Our network contains two separate model to fully regress and enhance the resulting poses. We utilize a newly proposed model whose name is Squeeze-and-Excitation Network as to construct our pose estimation network in order to estimate the corresponding pose from a color image; Then a model consisting of several blocks of fully connected networks and a novel semantic graph convolutional networks featuring self-supervision to reconstruct 3D human pose. We demonstrate the effectiveness of our approach on standard datasets for benchmark where we achieved comparable results to some recent state-of-the-art methods existed.

Key words - Pose estimation; 3D human pose regression; single view; convolutional network; deep learning.

1. Đặt vấn đề

Những năm gần đây, mô phỏng dáng người trên không gian ba chiều từ ảnh màu đơn đang là một hướng nghiên cứu nhận được nhiều sự chú ý quan tâm đặc biệt, bởi tiềm năng ứng dụng của nó vào thực tiễn đời sống phong phú của con người, ví dụ như cơ sinh học, hệ thống giám sát, thực tế ảo và thực tế ảo tăng cường [1], [2]. Tuy nhiên, những phương pháp phổ thông hiện nay dùng để thu thập dữ liệu mô hình người trong không gian ba chiều vẫn còn thiếu sự linh hoạt, cũng như khá tốn kém về mặt chi phí trong việc thực hiện, dẫn đến cần một phương pháp đơn giản hơn để có thể thực hiện việc dựng hình mà ít phải thêm vào các phương thức phức tạp đã có sẵn ở cách dựng mô hình hai chiều. Ngoài ra, tuy lĩnh vực này còn rất nhiều triển vọng, nhưng vẫn còn không ít khó khăn để thực hiện do sự hạn chế từ cơ sở dữ liệu vẫn còn khiếm khuyết trong việc miêu tả hình thể, sự khác biệt giữa các góc máy tới đối tượng, và những ràng buộc về không gian.

Những thành tựu gần đây của mạng chập nơ-ron (CNN hay ConvNet) [3] đã giúp cho việc phát triển các mô hình mô phỏng dáng người trong không gian ba chiều đạt những bước tiến mới. Có thể kể đến như, các phương pháp nâng từ dáng người hai chiều cộng với các kĩ thuật học sâu đã giúp cho mô phỏng dáng người ba chiều đạt đến kết quả tham chiếu nhờ kết hợp các phương thức trên (ví dụ như Convolutional Pose Machine (CPM) [4], Stacked Hourglass Networks [5]). Tuy nhiên, phần lớn các mô hình vẫn còn dựa vào đầu vào là dáng người hai chiều có sẵn từ các mô hình dựng dáng người hai chiều [6], [7], hoặc là chỉ tập trung vào các phương thức ánh xạ từ 2D sang 3D [8] [9]. Dù cho kết quả là rất tốt so với thời điểm đó, các phương thức trên vẫn còn bị hạn chế bởi vẫn còn rất nặng về mặt tính toán vì sử dụng các mạng phức tạp, dẫn đến việc áp dụng vào nhiều điều kiện môi trường trong thực tế vẫn còn chưa đạt được kì vọng [10].

Trong bài báo này, nhóm tác giả đề xuất một mô hình dựa trên heatmap và hồi quy các vị trí của các điểm khớp (joint positions) để dựng lại thành mô hình khung xương trong không gian ba chiều. Phương pháp này sử dụng một ảnh đơn từ đầu vào và biến đổi nó thành dáng người hai chiều (2D keypoints/2D pose) để rồi từ đó dựng nó thành dáng người ba chiều (3D keypoints/3D pose).

2. Các nghiên cứu liên quan

Mô phỏng dáng người có thể chia làm hai phương thức tiếp cận: Phương pháp tạo từ mô hình và phương pháp phân loại.

- Mô hình theo cấu trúc ảnh (PSM) là một trong những mô hình tái tạo cho mô phỏng dáng người hai chiều khá phổ biến hiện nay. PSM chiếu hình ảnh người thành mô hình các khớp nối. Mô hình này thường chia làm hai phần: Một là biểu diễn các điểm khớp trên cơ thể, hai là

[1] The University of Danang - University of Science and Technology (Pham Le Minh Hoang, Le Thi Kim Oanh)

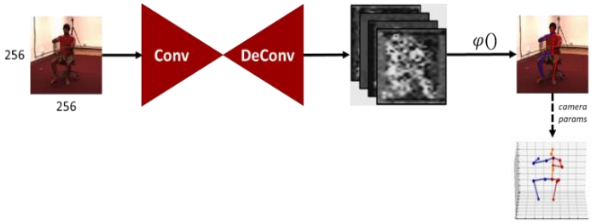
mỗi quan hệ giữa các điểm đó. Bởi vì chiều dài của hông trên không gian hai chiều là không cố định, một tổ hợp các mô hình được đề xuất để dựng từng phần. Mối quan hệ trong không gian giữa các điểm trong không gian ba chiều dễ biểu diễn hơn đối với mô phỏng dáng người ba chiều, khi mà chiều dài của hông là cố định cho mỗi đối tượng. Burenius và cộng sự [11] đề xuất áp dụng PSM vào mô phỏng dáng người trong không gian ba chiều bằng cách ước lượng xấp xỉ chiều sâu trong không gian. Tuy nhiên, dáng người trong không gian lũy thừa theo mũ 3, dẫn đến độ phức tạp quá lớn.

- Phương pháp phân loại xem việc mô phỏng dáng người như là một bài toán hồi quy. Sau khi trích xuất đặt trưng từ ảnh, một sơ đồ được học từ không gian đặc trưng thành không gian dáng người. Bởi vì tính chất mô hình khung xương, quan hệ vị trí của từng khớp là tương đối quan trọng. Để tính toán từng biến độc lập giữa các biến đầu ra, Ionescu và cộng sự [12] đề xuất dùng Support Vector Machine (SVM) để học sơ đồ từ các đặc trưng từng phần thành vị trí các khớp.
- Với tiếp cận theo phương thức học sâu, thay vì phải giải quyết các vấn đề về các điểm trên hình thể bằng cách thủ công ở từng điểm một, một phương pháp trực tiếp hơn là "nhúng" cả mô hình vào một hàm ánh xạ và học cách biểu diễn. Trong trường hợp này, mô hình cần phải học được đặc điểm chung của dáng người trong dữ liệu, dẫn đến cần một bộ dữ liệu lớn để học.

3. Mô hình và phương pháp

Mô hình của nhóm được thừa hưởng ý tưởng thiết kế của Xiao và cộng sự [13], Sun và cộng sự [14]. Để đi vào chi tiết vào mặt thiết kế mô hình, nhóm tác giả chia mô hình tổng thành hai phần. Phần đầu liên quan chính đến sử dụng mô hình mạng chập học sâu để thu được heatmap 3D của từng điểm khớp trong tọa độ ảnh và chiều sâu với xương chậu là gốc tọa độ. Phần sau của mạng nhận đầu ra của phần trước làm đầu vào và đưa tiếp vào các lớp kết nối đầy đủ và mạng chập đồ thị để "nâng" và tăng cường độ chính xác về chiều sâu. Kết quả cuối cùng ta thu được mô phỏng dáng người trong không gian ba chiều hoàn chỉnh.

3.1. Mạng mô phỏng dạng người – poSEnet

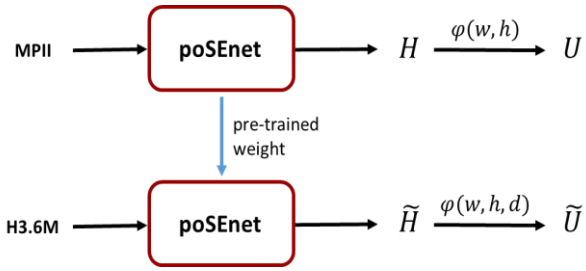


Hình 1. *Mô hình để xuất để mô phỏng dáng người*

Trong phần này được miêu tả trong Hình 1, bằng cách sử dụng mô hình đã huấn luyện trên bộ dữ liệu ImageNet đã có trước đây, nhóm tác giả chỉnh sửa lại thành mô hình để mô phỏng dáng người bằng phương pháp transfer learning. Mạng lưới này bao gồm mô hình mạng chập sâu để trích xuất đặc trưng của ảnh từ đầu vào, và cho vào tiếp một mạng khử chập (deconvolutional network) để upsample thu được đầu ra như mong muốn là các sơ đồ đặc trưng [13], [14], [15]. Mặc định, ba lớp mạng chập chuyển vị để khử chập, được sử dụng với batch normalization [16] và hàm ReLU [17]. Mỗi lớp có 256 bộ lọc với kích thước nhân kernel 4x4 và stride là 2. Một lớp mạng chập kích thước 1x1 để tạo sơ đồ đặc trưng được dự đoán cho tất cả các điểm khớp. từ đó thu được sơ đồ đặc trưng với kích thước là 64x64x64xJ với J là số khớp nối trên mô hình khung xương để biểu diễn dáng người.

Trong mạng chập, thay vì sử dụng trực tiếp ResNet để giảm kích thước đầu vào, nhóm tác giả đề xuất sử dụng mạng Squeeze-and-Excitation Networks (SE) [18]. Mô hình gốc khi chạy trên ImageNet [3] cho kết quả vượt trội hơn ResNet-50 0,86% và tiệm cận ResNet-101 với số tham số ít hơn rất nhiều làm giảm hơn một nửa chi phí tính toán. Mạng sử dụng kiến trúc "ép-giãn" (SE) để nén thông tin từ ảnh đầu vào và giải nén trở lại thành sơ đồ đặc trưng.

Quy trình huấn luyện được thể hiện trong Hình 2. Đầu tiên mạng sẽ được huấn luyện trước trên bộ dữ liệu MPII [19]. Ảnh được đưa vào mạng mô phỏng dáng người để thu được heatmap $H \in \mathbb{R}^{w \times h}$, với w, h là kích thước sau khi khử chập. Bằng cách áp dụng hàm soft-argmax đề xuất bởi Sun [14], kết quả thu được là dáng người trong không gian hai chiều. Lí do để sử dụng MPII làm tiền huấn luyện là vì để cho mạng học trước một số thông tin về mô phỏng, giúp giảm thời gian và tài nguyên tính toán khi đưa bộ dữ liệu dáng người trong không gian ba chiều vào học. Bước kế tiếp chỉ sử dụng H3.6M [20] để học cấu trúc dáng người ba chiều từ ảnh đầu vào. Cùng kĩ thuật được áp dụng khi tiền huấn luyện với MPII, nhưng heatmap thu được sẽ là $H \in \mathbb{R}^{w \times h \times d}$, với w, h vẫn là kích thước sau khi khử chập, d là chiều sâu ước lượng được định nghĩa như một siêu tham số, sau đó sử dụng hàm soft-argmax để thu được dáng người gồm ba chiều (không cùng một hệ tọa độ).

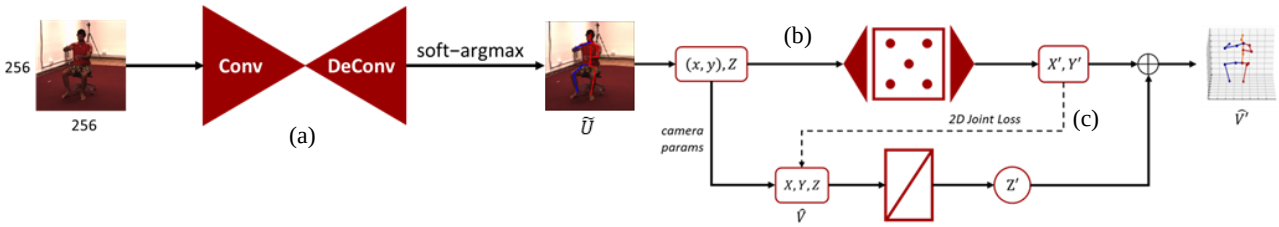


Hình 2. *Các bước training mô hình*

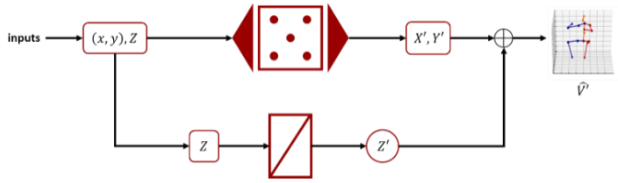
Tuy nhiên, về mặt chính xác mà nói, thì đầu ra của mô hình này không phải là ba chiều hoàn chỉnh trong không gian, mà là bao gồm dáng người hai chiều trên tọa độ ảnh (x_{img}, y_{img} với x và y là tọa độ trong không gian ảnh), và tọa độ Z là chiều sâu của các điểm khớp trong không gian với gốc tọa độ là khung xương chậu Z_{pelvis}. Lí do vì, khi mô hình học các điểm khớp từ dữ liệu đầu vào, nó không thể học trực tiếp từ một ảnh đơn hai chiều không hề có dữ liệu về chiều sâu trong ảnh. Vì thế, đây là kết quả nội suy từ mô hình từ giá trị dữ liệu thật của bộ dữ liệu.

3.2. Dựng dáng người trong không gian ba chiều kết hợp với học tự giám sát

Trong mô hình đề xuất ở phần này (Hình 4), nhóm tác giả kết hợp sử dụng hai mô hình nhỏ hơn. Phần trên được gọi là mạng chập đồ thị SemGCN đề xuất bởi Long Zhao [21]; Phần dưới là mạng tuyến tính đề xuất bởi Martinez [22]. Một thành phần tự học giám sát được thêm vào để cập nhập sai số.



Hình 3. *Cấu trúc của mô hình mô phỏng dáng người trong không gian ba chiều có kết hợp học tự giám sát. Mô hình được chia làm hai thành phần chính: (a) Mô-đun mô phỏng dáng người dùng để tái tạo mô hình dáng người trong không gian hai chiều và học độ sâu của mỗi điểm khớp so với tọa độ gốc là hông/xương chậu; (b) Mô-đun hồi quy dùng để "nâng" các tọa độ của dáng người trong không gian hai chiều ở ảnh (x_{img}, y_{img}) sang ba chiều, bao gồm hai nhánh mô hình để hồi quy và tăng cường độ chính xác; (c) Một nhánh học tự giám sát được thêm vào để mô hình học được cách "nâng" tọa độ ảnh sang tọa độ X, Y trong không gian ba chiều một cách chính xác hơn. Bên cạnh đó, mô hình tuyến tính ở nhánh dưới chỉ nhận tọa độ Z là đầu vào để tăng cường độ chính xác.*

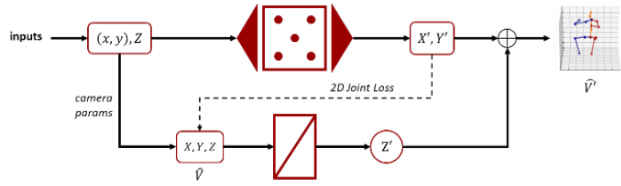


Hình 4. *Mô đun hồi quy*

Mạng chập đồ thị SemGCN. Trong nhánh trên của mô hình, nhóm tác giả điều chỉnh mạng này nhằm nâng các tọa độ khớp hai chiều vào vị trí chung trong không gian ba chiều. SemGCN học cách nắm bắt thông tin ngữ nghĩa như các mối liên hệ đỉnh cục bộ và toàn cục, vốn không được biển diễn rõ ràng trong đồ thị, có thể học được thông qua huấn luyện đầu cuối từ nhãn dữ liệu.

Mô hình tuyến tính tăng cường. Dựa trên một mạng nơ-ron nhiều lớp, sâu và đơn giản. Mạng này có 2 khối tính toán, gồm một số lớp tuyến tính nhất định có nối dư (residual connection) cùng với batch normalization [16], hàm ReLU [17] và các lớp dropout nhằm ánh xạ các nhiễu đầu vào từ đầu ra trước thành dáng người trong không gian ba chiều với độ tin cậy cao hơn. Thay vì "nâng" các điểm khớp trong không gian hai chiều, chúng tôi cho đầu vào với kích thước 3J (ba chiều) vào mạng này nhằm thu về các tọa độ của các khớp trong không gian ba chiều đã được tăng cường (với xương chậu làm gốc tọa độ) có kích thước cũng là 3J. Kích thước của mỗi lớp tuyến tính bên trong là 1024. Mạng này hưởng lợi từ nhiều đối với việc tối ưu hóa mạng nơ-ron sâu, thường xuất hiện trong các bài báo gần đây về cải thiện những mạng CNN trong học sâu.

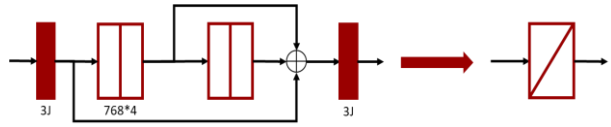
Hoạt động như một mô đun để hồi quy, mô hình ở phần này lấy đầu vào chia dữ liệu hai chiều từ ảnh và tọa độ Z theo trục tọa độ không gian ba chiều với xương chậu làm gốc tọa độ. Tuy nhiên, khi đưa trực tiếp đầu vào trên vào SemGCN [21] để dựng dáng người hai chiều thành ba chiều, việc nội suy của mô hình sẽ phải đối mặt với việc thiếu thông tin về chiều sâu trong thực tế. Ngược lại, nếu chỉ để mô hình tuyến tính làm mô đun hồi quy như cách đề xuất ở bài báo gốc [22], nó lại có thể nội suy tọa độ Z của các khớp tốt hơn là sử dụng mạng chập đồ thị, nhưng lại giảm đi độ chính xác của việc dự đoán X và Y rất nhiều. Với những vấn đề vừa nêu, thông thường thì cách tăng độ chính xác chỉ đơn giản là là tăng số lớp ẩn, nhưng đấy cũng đồng thời làm tăng kích thước của mô hình với cấp số nhân. Ví dụ như ở [23], việc tăng số lớp ẩn từ 1024 lên 2048 đã tăng số tham số của mô hình từ 4 triệu lên 7 triệu, nhưng độ chính xác tăng lên lại không hề tương ứng. Do đó, bằng cách tận dụng các ưu điểm của từng mô hình dựa trên những đặc điểm của nó, nhóm tác giả đề xuất mô hình ở trong Hình 5 như sau:



Hình 5. *Huấn luyện cho mô đun hồi quy*

Đối với nhánh trên sử dụng mạng chập đồ thị SemGCN, nhóm tác giả lấy tất cả các kích thước đầu vào để huận luyện mô hình, nhưng chỉ nhận X' và Y' làm giá trị đầu ra. Xây dựng lại cách biểu diễn các điểm khớp để dựng đồ thị, bằng cách nhóm mô hình 17 khớp thành 9 nhóm phần trên và dưới nhằm xác định ma trận kề (adjacency matrix) đã đề cập ở trên. Nhóm tác giả nhận ra đối với mô hình này, tăng số lượng lớp ẩn từ 128 lên 256 cũng làm giảm sai số các điểm khớp ở mức vừa phải.

Đối với mô hình tuyến tính ở nhánh dưới, kiến trúc tổng thể vẫn được lấy cảm hứng từ mô hình ban đầu. Bằng cách sử dụng các thông số của máy ảnh để tái tạo lại dáng người trong ảnh thành dáng người trong không gian ba chiều, nhóm tác giả có thể huấn luyện nhánh dưới như một mô đun tăng cường thuần túy. Do đó, chỉnh sửa lớp đầu vào của nhánh này để nhận dữ liệu đầu vào là ba chiều và cho nó học để tăng cường độ chính xác của chiều sâu trong không gian. Khi quan sát đặc điểm trong quá trình huấn luyện ở nhiễu đầu ra do sai số giới hạn của mô hình tuyến tính, chỉ lấy tọa độ Z ở đầu ra. Để phóng được từ kích thước đầu vào không phải nhị phân sang các lớp ẩn, điều chỉnh lại kích thước của các lớp tuyến tính xuống còn 768. Do tỉ lệ giữa đầu vào và lớp tuyến tính nhỏ hơn so với mô hình ban đầu, vốn có thể dẫn đến các sai số lớn hơn, nhóm tác giả tăng số tầng và thêm các một vài đoạn nối dư nhằm tạo điều kiện thuận lợi cho luồng thông tin giữa các lớp (Hình 6). Hơn nữa, chúng cũng giúp cải thiện hiệu suất và giảm thời gian huấn luyện.



Hình 6. *Mô hình tuyến tính đề xuất và kí hiệu thu gọn*

Cuối cùng, nhóm tác giả nối đầu ra của mỗi nhánh để thu được dáng người ba chiều hoàn thiện. Ở giữa hai nhánh,

để tăng cường tính hiệu quả cho việc sử dụng luồng thông tin giữa các nhánh, thêm vào bước học tự giám sát để tính sai số giữa nhánh hồi quy phía trên và nhánh sử dụng thông số máy ảnh để dựng ở phía dưới, và chỉ tính sai số giữa X, Y với X', Y' như minh họa ở Hình 5. Trong phần này, các mô hình được huấn luyện độc lập.

Tiền xử lí dữ liệu. Trước khi cho các sơ đồ đặc trưng vào hàm soft-argmax để thu được xác suất của các điểm khớp trên ảnh, chúng tôi dùng chuẩn hóa tuyến tính để giảm kích thước về khoảng [-1, 1] dựa trên công thức:

$$(x, y, Z)' = \frac{(x,y,Z)}{64} - 0{,}5 \quad (1)$$

với (x, y) là tọa độ trong không gian ảnh; Z là tọa độ trong không gian ba chiều với xương chậu là gốc tọa độ. Để tiện trong việc dựng lại dáng người trong không gian ba chiều bằng thông số máy ảnh, đầu ra sau khi đưa vào hàm soft-argmax được tính ngược lại về [0, 255] và [-128, 127] tuần tự cho tọa độ (x, y) và Z. Để tính MPJPE, tọa độ của các điểm khớp trong không gian ba chiều của máy ảnh từ dữ liệu đánh nhãn cũng chuyển thành tọa độ trong không gian ba chiều với xương chậu làm gốc tọa độ, và các trục tọa độ sẽ chạy trong khoảng [-1000; 1000] milimét.

Ở phần sau, chuẩn hóa tuyến tính được sử dụng để nhằm huấn luyện nhanh hơn và kết quả hội tụ chính xác hơn. Công thức được sử dụng là:

$$s' = \frac{s}{ImageSize} \quad (2)$$

4. Thí nghiệm và đánh giá kết quả

4.1. Bộ dữ liệu

Trong nghiên cứu này, nhóm tác giả tiến hành trên bộ dữ liệu Human3.6M (H3.6M), dữ liệu lớn nhất cho việc đánh giá kết quả mô phỏng dáng người trong không gian ba chiều [20]. Dữ liệu này chứa 3,6 triệu ảnh từ 11 người khác nhau (6 nam và 5 nữ), thực hiện 15 hành động thường ngày như ăn, đứng, đi bộ, chụp ảnh, cũng như các hoạt động khác thu được từ 4 góc máy khác nhau cùng lúc với kích thước ảnh là 1000x1000 pixel.

4.2. Phương thức đánh giá

Nhóm tác giả đi theo các phương thức đánh giá tiêu chuẩn khi sử dụng cả 4 góc máy từ đối tượng 1, 5, 6, 7, 8 để huấn luyện mô hình, và cũng dùng cả 4 góc máy ở đối tượng 9 và 11 để kiểm thử. Trong cả quá trình huấn luyện và đánh giá, tần số lấy mẫu là 5 Hz. Độ chính xác được đánh giá bằng MPJPE (viết tắt từ tiếng ảnh của "giá trị sai số trung bình của các khớp"), để tính toán trên Phương thức đánh giá #1 (PTĐG #1); và "sai số với hệ PA" tức là sắp xếp lại dáng người trong không gian ba chiều và ground truth bằng cách sử dụng Procrustes Analysis [24] và đánh giá ở Phương thức đánh giá #2 (PTĐG #2). Ở các PTĐG, giá trị mong muốn thu được càng nhỏ càng tốt.

4.3. Kết quả

Kết quả thu từ heatmap (nửa đầu). Để tính MPJPE ở nửa đầu của mô hình tổng, nhóm tác giả phải sử dụng dữ liệu thông số ảnh từ máy ảnh để tính toán và dựng lại tọa độ không gian ba chiều nơi mà bộ dữ liệu này được thu thập. Trong Bảng 1, kết quả thu được được so sánh với các phương pháp từ các nghiên cứu khác.

Bảng 1. *Kết quả và so sánh giữa mô hình thuần nửa đầu và các phương thức khác*

	PTĐG #1	PTĐG #2	Số lượng tham số (M)
Muhammed và cộng sự. [15] – ResNet-50	51,83 mm	45,04 mm	34,291
Sun và cộng sự [14] – ResNet-50	49,60 mm	40,60 mm	34,291
Lie và cộng sự [23] – ResNeXt-50	50,44 mm	38,93 mm	33,763
Của nhóm TG – SE-ResNeXt-50	49,28 mm	43,01 mm	36,281

ResNeXt tự thân là một kiến trúc mạng dạng mô đun dành cho các tác vụ thị giác máy tính. Nó được xây dựng từ kiến trúc nhiều nhánh và đồng nhất chỉ với một lượng ít siêu tham số thiết lập. Khi đi cùng với khối SE, độ sâu và độ chính xác của mô hình học được được cải thiện đáng kể. Ngoài ra, lợi ích từ việc các đặc trưng được trích xuất có thể được tăng cường dần nhờ các khối SE. Trong bài báo này, khi sử dụng ResNeXt-50 với khối SE cho ra kết quả tốt hơn so với các kết quả từ các bài báo khác.

Kết quả thu được từ hồi quy (nửa sau). Để huấn luyện trong phần này, nhóm tác giả sử dụng đầu vào từ đầu ra của mạng trước. Sau khi huấn luyện hai nhánh của mô hình này riêng biệt, việc đánh giá được thực hiện lại một lần nữa khi kết nối hai nhánh lại với nhau. Kết quả so sánh được ghi lại ở Bảng 2.

Bảng 2. *Kết quả so sánh với các phương pháp khác cũng sử dụng mạng hồi quy (nửa sau)*

	PTĐG #1	PTĐG #2	Số lượng tham số (M)
Martinez và cộng sự [22] (sử dụng cùng đầu vào)	51,03 mm	38,78 mm	4,29
Zhao và cộng sự [21] (sử dụng cùng đầu vào)	49,92 mm	38,66 mm	0,43
Lie và cộng sự [23]	51,18 mm	38,89 mm	17,00
Pavlakos và cộng sự [28] (*)	46,80 mm	36,50 mm	16,95
Của nhóm TG	47,34 mm	37,26 mm	6,53

(*) Phương pháp sử dụng đầu vào là dạng chuỗi

Nếu xét trường hợp các thông số để dựng lại không gian ba chiều được cung cấp, thì phần sau gần như hoạt động với tư cách là mô đun tăng cường độ chính xác của kết quả. Tuy nhiên, để làm được điều này thì cần có thông số từ máy ảnh sử dụng để ghi ảnh hoặc video để dựng từ dáng người trong không gian ảnh hai chiều sang không gian ba chiều. Các thông số ảnh thì có thể trích xuất từ EXIF của ảnh; còn khoảng cách từ ống kính để người hoặc vật thể thì chỉ có thể tính thông qua công thức với chiều cao của vật thể phải biết trước. Vì thế nếu trường hợp là một đám đông đa dạng về đặc điểm chiều cao khác nhau, ước lượng khoảng cách sẽ dễ gây ra sai số.

Nếu hoạt động như mô đun hồi quy, nó có thể dùng như một mô đun rời kết hợp với các mô hình mô phỏng dáng người trong không gian ảnh hai chiều đạt kết quả tham

chiếu. Trong bài báo, nhóm tác giả tiến hành đánh giá trên mô hình đề xuất ở nửa đầu, và cũng sử dụng cùng tần số lấy mẫu trên bộ dữ liệu Human3.6M. Như đã đề cập, khi không có các thông số máy ảnh, mô hình của nhóm tác giả đề xuất vẫn có thể hồi quy và tăng cường độ chính xác của kết quả mà không bị tăng sai số.

So với các mô hình khác, phương pháp hồi quy nhóm tác giả đề xuất có số các tham số của mô hình vừa phải hơn mà vẫn đạt được kết quả tốt như mong đợi. Tuy nhiên, vì nó có tính liên kết chặt chẽ với kết quả từ mô-đun dựng dáng người trong không gian hai chiều, so sánh này chỉ mang tính tương đối. Tùy thuộc vào nhu cầu giữa độ chính xác và tốc độ xử lí dựa trên thiết bị phần cứng sẵn có, mô hình và tham số có thể hiệu chỉnh để phù hợp. Tuy nhiên, nhóm tác giả khuyến khích khi hiệu chỉnh cần thực hiện kiểm thử và huấn luyện lại để tránh tình trạng bị overfit.

Phân tích thành phần. Nhóm tác giả cũng thử nghiệm trên từng phần của mô hình dựa trên bộ dữ liệu Human3.6M trên cả 2 phương thức đánh giá để đánh giá từng thành phần của mô hình. Các kết quả được ghi lại ở Bảng 3 khi thay đổi các từng thành phần.

Để hiểu thêm về từng thành phần, nhóm tác giả bắt đầu với kết quả từ mô đun mô phỏng dáng người trong không gian hai chiều và dựng thành dáng người trong không gian ba chiều qua thông số của máy ảnh để tính MPJPE. Nếu sử dụng mô hình tuyến tính đơn giản đề xuất ở [22] như là một mô đun hồi quy với 1024 lớp ẩn, kết quả dường như kém chính xác hơn khi nó thiếu thông tin để dựng thành dáng người trong không gian ba chiều từ hai chiều vì thiếu thông tin trong chiều sâu trong không gian. Tương tự cũng xảy ra với SemGCN [21], nhưng sai số nhỏ hơn do tính phức tạp hơn của mô hình. Sau khi gắn cả hai mô hình vào với nhau thành một mô đun với nhánh hồi quy ở trên và nhánh tăng cường ở dưới, sai số đã giảm đi đáng kể. Sau khi chuyển đổi mô hình tuyến tính thuần thành mô hình nhóm tác giả đề xuất, sai số đã giảm từ 51,03mm xuống 47,43mm.

Bảng 3. *Phân tích từng thành phần*

Phương pháp	PTĐG #1	PTĐG #2
poSEnet (dựng lại từ các thông số của máy ảnh)	49,28 mm	43,01 mm
Linear regression [22] (chỉ mô hình ở nhánh dưới)	51,03 mm	38,78 mm
SemGCN regression [21] (chỉ mô hình ở nhánh trên)	49,50 mm	38,21 mm
Full regression module (học tự giám sát)	48,22 mm	37,99 mm
Full regression module (SemGCN 128 -> 256)	47,89 mm	37,85 mm
Full regression module (đơn giản -> mô hình tuyến tính được đề xuất)	47,34 mm	37,26 mm

4.4. So sánh và đối chiếu

Ở Bảng 4, nhóm tác giả thực hiện so sánh với các phương pháp đạt kết quả tốt nhất hiện tại sử dụng một góc máy để dựng lại dáng người trong không gian ba chiều trong những năm gần đây. Để mang tính nhất quán, tất cả đều cùng được so sánh cùng hệ quy chiếu trên cả 2 phương thức đánh giá. So với các phương pháp khác, mô hình của nhóm tác giả đạt được kết quả tương đối tốt so với các mô hình đạt chuẩn tham chiếu hiện tại. Ở một số hành động, mô hình còn cho kết quả tốt hơn. Điều đó cho thấy, tính hiệu quả của mô hình đề xuất khi đạt được kết quả tương đối khả quan so với các mô hình tốt nhất hiện tại.

Bảng 4. *So sánh giữa các phương pháp theo Phương thức đánh giá#1 trên bộ dữ liệu Human 3.6M*

PTĐG #1	Direction	Discuss	Eat	Great	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Zhou và cộng sự (ICCV'17) [10]	54,8	60,7	58,2	71,4	62,0	65,5	53,8	55,6	75,2	111,6	64,1	66,0	51,4	63,2	55,3	64,9
Martinez và cộng sự (ICCV'17) [22]	51,8	56,2	58,1	59,0	69,5	78,4	55,2	58,1	74,0	94,6	62,3	59,1	65,1	49,5	52,4	62,9
Fang và cộng sự (AAAI'18) [27]	50,1	54,3	57,0	57,1	66,6	73,3	53,4	55,7	72,8	88,6	60,3	57,7	62,7	47,5	50,6	60,4
Pavlakos và cộng sự (CVPR'18) [28]	48,5	55,4	54,4	52,0	59,4	65,3	49,9	52,9	65,8	71,1	56,6	52,9	60,9	44,7	47,8	56,2
Sun và cộng sự (ECCV'18) [14]	46,5	48,1	49,9	51,1	47,3	43,2	45,9	57,0	77,6	47,9	54,9	46,9	37,1	49,8	41,2	49,8
Zhao và cộng sự (CVPR'19) [21]	47,3	60,7	51,4	60,5	61,1	49,9	47,3	68,1	86,2	55,0	67,8	61,0	42,1	60,6	45,3	57,6
Chen và cộng sự (CVPR'19) [26]	**41,1**	**44,2**	44,9	45,9	**46,5**	**39,3**	41,6	54,8	73,2	**46,2**	48,7	**42,1**	**35,8**	46,6	38,5	**46,3**
Pavllo và cộng sự (CVPR'19) [25] (*)	45,2	46,7	43,3	**45,6**	48,1	55,1	44,6	**44,3**	**57,3**	65,8	**47,1**	44,0	49,0	**32,8**	**33,9**	46,8
Wen-Nung Lie và cộng sự (2019) [23]	43,2	49,1	45,7	64,4	49,8	54,8	42,9	45,5	58,4	76,3	47,5	58,8	50,0	38,0	40,3	51,2
Của nhóm TG	43,5	47,2	**42,3**	46,2	47,7	41,0	**41,3**	55,5	63,8	47,0	53,1	47,7	36,8	47,6	40,0	47,3

5. Kết luận

Với phương pháp mà nhóm tác giả đề xuất cho việc mô phỏng dáng người trong không gian ba chiều, đã đạt được các kết quả như sau:

- Phương thức đánh giá #1: 47,34 mm;
- Phương thức đánh giá #2: 37,26 mm.

Kết quả cho thấy, tính hiệu quả và tính linh hoạt của mô hình được đề ra. Trong nghiên cứu này, nhóm tác giả cũng đã chỉ ra rằng với một mô hình học sâu đơn giản, hiệu quả, kết hợp với sử dụng mạng chập đồ thị cùng với đó kết hợp học tự giám sát, đã cho ra một kết quả tương đối chính xác và có thể so sánh với các phương pháp đạt chuẩn tham chiếu. Tính đơn giản trong mô hình của bài báo đề xuất mở ra các

hướng nghiên cứu mới trong tương lai. Ví dụ, nhờ sự linh hoạt của mô hình, nó có thể kết hợp được với một số các mô hình đã có để bổ trợ cho kết quả cuối cùng thu được; Hoặc có thể tích hợp vào một trong các mô đun của các phương thức sử dụng nhiều góc máy ảnh (các phương thức này thường sẽ đạt kết quả tốt và tốn ít chi phí tính toán hơn).

Cho đến hiện tại, ứng dụng của các phương pháp sử dụng mô phỏng dáng người trong không gian ba chiều vẫn đang còn nhiều tiềm năng chưa được khai phá hết. Trên thực tế, nó thường được dùng như một bài toán trung gian trong một bài toán lớn hơn trong lĩnh vực thị giác máy tính (ví dụ như nhận diện hành động). Nếu được nghiên cứu và ứng dụng sâu hơn vào các bài toán nhận diện và phân tích hành động, cử chỉ áp dụng trong các thiết bị giám sát, nó có thể mở ra thêm khả năng ứng dụng trong bài toán quản lí chất lượng nhân sự với tiềm năng từ dữ liệu ba chiều.

Lời cảm ơn: Bài báo này được tài trợ bởi Trường Đại học Bách khoa – Đại học Đà Nẵng với đề tài có mã số: T2021-02-42.

TÀI LIỆU THAM KHẢO

[1] Connolly, I., Palmer, M., Barton, H., & Kirwan, *An Introduction to Cyberpsychology*, Routledge, 2016.

[2] C. Held, J. Krumm, P. Markel, and R. P. Schenke, "Intelligent video surveillance", *Computer*, Vol. 45, 2012, 83–84.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems*, Vol. 25, 2012, 1097-1105.

[4] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional Pose Machines", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 4724-4732.

[5] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation", *Computer Vision – ECCV 2016*, 2016, 483-499.

[6] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 7025-7034.

[7] D. Tome, C. Russell, and L. Agapito, "Lifting from the Deep: Convolutional 3D pose estimation from a single image", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 2500-2509.

[8] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image", *Computer Vision and Image Understanding*, Vol. 172, 2018, 37-49.

[9] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: a weakly-supervised approach", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, 398-407.

[10] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3D human pose estimation in the wild by adversarial learning", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 5255-5264.

[11] Magnus Burenius, Josephine Sullivan, Stefan Carlsson, "3D Pictorial Structures for Multiple View Articulated Pose Estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, 3618-3625.

[12] Catalin Ionescu, Liefeng Bo, Cristian Sminchisescu, "Structural SVM for visual localization and continuous state estimation", *Proceedings of 12th International Conference on Computer Vision (ICCV)*, 2009, 1157-1164.

[13] Bin Xiao, Haiping Wu, and Yichen Wei, "Simple Baselines for Human Pose Estimation and Tracking", *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 466-481.

[14] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei, "Integral human pose regression", *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 529-545.

[15] Muhammed Kocabas, Salih Karagoz, Emre Akbas, "Self-Supervised Learning of 3D Human Pose using Multi-view Geometry", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 1077-1086.

[16] Sergey Ioffe, Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal co-variate shift", *Proceedings of the 32nd International Conference on Machine Learning*, 2015, 448-456.

[17] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng., "Rectifier non-linearities improve neural network acoustic models", *Proceedings of the International Conference on Machine Learning*, Vol. 28, 2013, 3-9.

[18] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu, "Squeeze-and-Excitation Networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 7132-7141.

[19] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele, "2D human pose estimation: New benchmark and state of the art analysis", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, 3686-3693.

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, 2014, 1325-1339.

[21] Long Zhao, Xi Peng, Yu Tian. Mubbasir Kapadia, Dimitris N. Metaxas, "Semantic Graph Convolutional Networks for 3D Human Pose Regression", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 3425-3435.

[22] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little, "A Simple yet Effective Baseline for 3D Human Pose Estimation", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, 2640-2649.

[23] Wen-Nung Lie, Lung-Sheng Shih, "3D Human Skeleton Estimation Based on 3D Heatmaps Generation and Regression by Deep Learning Techniques", *National Chung Cheng University Online Library*, 2019, https://hdl.handle.net/11296/5z969r, 14/02/2020.

[24] J. C. Gower, "Generalized procrustes analysis", *Psychometrika*, 1975, 33–51.

[25] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli, "3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training", *Proceedings of IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, 7753-7762.

[26] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin, "Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 10895-10904.

[27] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu, "Learning pose grammar to encode humanbody configuration for 3D pose estimation", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, 2018, 6821–6828.

[28] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. "Ordinal depth supervision for 3D human pose estimation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 7307-7316.