ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.1729

LiteNetPose: A lightweight neural network for human pose estimation using attention modules

Jorge L. Charco, Dr.[1], Angélica Cruz-Chóez, Msc.[1], Angela Yanza-Montalván, Msc.[1],
Johanna Zumba-Gamboa, Msc.[1], María Galarza-Soledispa, Msc.[1]
[1]Universidad de Guayaquil, Ecuador, jorge.charcoa@ug.edu.ec, angelica.cruzc@ug.edu.ec, angela.yanzam@ug.edu.ec,
johanna.zumbag@ug.edu.ec, maria.galarzas@ug.edu.ec

***Abstract*—This paper presents the usage of attention modules to tackle the challenging problem of the self-occlusion cases in human pose estimation problem. The proposed approach first obtains the relevant features of the human body joints of a set of images using ResNet-50 architecture (just 5.5% of the 25.6M parameters available are considered) as backbone, which are captured from different views at the same time. Then, a Bone position encoding is proposed to obtain the information about position and orientation of body bone, mainly, those bones whose body joints have more probability to be occluded due to the natural human body pose. These obtained results together with the obtained relevant features of the human body joints using ResNet-50, are used as input to the attention module. Basically, the body joints from a given view are used to enhance poorly estimated joints from another view due to the self-occlusion cases. Experimental results and comparisons with the state-of-the-art approaches on Human3.6m dataset are presented showing improvements in the accuracy of body joints estimations.**

***Keywords*—human pose estimation, attention modules, multi-view environments, neural networks.**

I. Introduction

The Human Pose Estimation (HPE) problem has been studied during the last decade. The main idea consists by first detecting each body joint such as elbow, knee, shoulder, etc, and then connecting them to get the human body skeleton. Some approaches have been proposed in the state-of-art such to tackle HPE problem such as OpenPose [1], DeepPose [2], Stacked Hourglass Networks [3], among others. These approaches are robust when all body joints are visible with respect to the position and orientation of camera while capturing images of the scene. However, the body joints are occluded regularly due to the natural human body pose while different activities are done, including certain objects (e.g., cars, bicycles, even other pedestrians) in the scene, which can occluded the body joints partially, being a challenging problem in monocular vision systems. Different applications require high accuracy of HPE to develop different solutions such as human action recognition, healthcare, augmented reality, just to mention a few.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI)
DO NOT REMOVE.

During last years, convolutional neural networks (CNNs) have been used for most of computer vision tasks (e.g., image enhancement, object detection and tracking, camera pose estimation, among others), getting better results with respect to classical approaches (e.g., [4], [5], [6]), including CNN architectures to solve HPE problems, which have showed appealing results (e.g., [7], [3], [8], [9], [10]). These architectures have been designed to tackle the HPE problem from monocular vision system scenarios, which use as input a set of images with single or multi-person. If multi-person input data are considered, the cost computational could be also increased due to the number of body joints of each subject to estimate in the image, and hence, also the inference time in real-time.

In recent years, the transformers, which have attention mechanisms, have been used to improve the results in natural language processing problems. On this basis, these mechanisms have been also applied in computer vision tasks such as object detection [11], [12], segmentation [13], [14], low-level vision task [15], including 3D human pose estimation [16], [17], showing appealing results due to the fact that they help to pay more attention to important areas and suppress other unnecessary information.

Although the results obtained for HPE problem are appealing, the occlusions of the body joints are not been completely solved, mainly, from monocular vision system. In order to overcome this problem, a new approach based on multi-view vision system is considered, which can capture the human body joints from different points of view at the same time. These multi-view approaches have been also used in occluded regions problems, such as camera pose, 3D-reconstruction or object detection [18]–[21]. Hence, the multi-view vision system is considered to tackle the human pose estimation problem, and thus improve the accuracy of occluded body joints. In details, the information obtained of human body joints from other cameras where these body joints are not occluded, are used to recover body joints occluded in one view at a snapshot in time. The contributions of this research are as follow:

- Develop a CNN architecture using attention mechanisms to tackle the human pose estimation problem considering multi-view vision systems.
- Show the importance of redundancy of information gener-

ated from other views to help to get more accuracy when the body joints are occluded.

In Section II previous works on human pose estimation are summarized; this is followed by Section III, which presents the proposed approach to obtain human body joints. The results of conducted experiments are reported in Section IV together with a detailed description of the used dataset and metrics; and finally, conclusions are given in Section V.

II. Related Work

Over time, many solutions have been implemented to solve the human pose estimation problems. CNN architectures are still used for the different computer vision tasks with appealing results. Nowadays, the transformer networks, specifically self-attention modules, are being used as complement to CNN architectures or even completely replacing them in computer vision tasks. On this address, the stacked Hourglass network, including it a polarized self-attention mechanism, have been proposed in [22]. In details, it is added before the second convolution of the basic residual block and the max pool down-sampling. However, the space and channel of the self-attention is required to maintain a high feature resolution. Other authors have introduced a self-attention mechanism in [23]. This mechanism combines long-term distance information and feature maps into original feature maps, generating an attention mask to re-weight original features, which force to the model focus more on non-local information. In details, to extract features from input images, a ResNet architecture is used, and then, the self-attention mechanism is considered to get long-range dependency between all body joints. The authors in [24] have proposed an attention refined network, which enhance multi-scale feature fusion for human pose estimation. Similar to previous work, channel and spatial attention mechanisms are considered, including a self-attention strategy that help to find long-range keypoints dependencies, which allowed to reinforce important features of input image. An approach based on regression method has been proposed in [25] to tackle human pose estimation problem. For this, the problem is formulated as a sequence prediction problem, which allow to attend important features to the target keypoints using attention mechanism in transformers. Three main components are used in the proposal: As first step is extract multi-level feature representations using a standard CNN backbone, and then, an encoder is considered to capture and fuse these multi-level features; and finally, a coarse-to-fine decoder to generate a sequence of keypoint coordinates.

Unlike the previous approaches, a purely transformer-based approach have been proposed in [26] to solve the human pose estimation in videos without using CNN architectures. The human body joints for each frame as well as the temporal correlations across frames are modeled using a spatial-temporal transformer structure. In details, the inputs to the spatial self-attention layers correspond a sequence of detected 2D poses, which help to generate a latent feature representation for each frame; and then, a temporal transformer is used to analyze the global dependencies between each spatial feature, which allow to obtain as output an accurate 3D human pose of the center frame. A similar work, other authors in [27] presented a Multi-level Attention Encoder-Decoder Network. The proposal included a Spatial-Temporal Encoder and a Kinematic Topology Decoder, which could model multi-level attentions. In order to extract the basic feature for each frame, a CNN backbone is used. The output of this backbone is sent to Spatial-Temporal Encoder. The spatial and temporal attention are learned from each block, which are processed from a series of cascaded blocks on Multi-Head self-attention. An unique linear regressor is used for each joint, which has been modeled using Kinematic Decoder. In order to calculate 3D joints and their 2D projection, these predicted parameters, including camera parameters, are utilized.

On the other hand, like in the single-view vision systems, transformer networks are being used from multi-view approaches to solve the human pose estimation problems, where these self-attention modules have showed appealing results. In [28], the authors have proposed to handle varying view numbers and video length without the need of camera calibration. In details, a backbone, which consist in a pre-trained 2D pose detector, is used to estimate the 2D pose from each image, and then, the predicted joints and its confidences are encoded into feature embedding for further 3D pose inference. These features embedding of each view are fused with a relative-attention block, which is able to relate each pair of views and reconstructs the features. The predict 3D human pose is obtained by adding these reconstructed features into a temporal fusing transformer. Other authors have proposed in [29], a multi-view pose transformer, which used a 2D pose detector proposed [30] by as first step to obtain high-resolution image features from multi-view inputs, then geometry-guide projective attention mechanism is used to fuse the multi-view information, instead of applying full attention to densely aggregate features across spaces and views. A novel RayConv operation is used to encode the camera rays into the multi-view feature representations to integrate multi-view positional information. In [31], the authors have proposed a similar approach to mentioned above. A transformer framework for multi-view 3D pose estimation is considered to improve the individual 2D predictors using information from different views. The low features are captured using the initial part of ResNet-50. The obtained features from the current view and neighboring view are fused, including an encoding 3D positional information to apply the concept of epipolar field into the transformer model. This allows to encode correspondences between pixels of different views.

III. Proposed Approach

The proposed approach consists to leverage the attention mechanisms, which was proposed by the original transformers in [32]. They allow to capture relevant features in the image to tackle the task of single human pose estimation. In details, The aim of the proposal is used an attention mechanism to capture long-range spatial relationship between the captured features

from a CNN backbone, which gets low-level image features, and then, use a head to predict the heatmaps of joints (see Fig. 1). According to [33], the ResNet-50 is used as backbone. For this, just 5.5% of the 25.6M parameters available of original architecture is considered for proposed approach. In details, given a set of pairs of images $I_i \in \mathbb{R}^{3\times H_i \times W_i}$, where $i \in 1, 2$ represents view 1 and view 2. The proposed backbone ß(.) obtains the low-levels features, as shown below:

$$X_i = ß(I_i) \in \mathbb{R}^{d\times H\times W}, \tag{1}$$

where the number of channels is denoted as d, and the height and width of the feature map are H and W, respectively.

2D spatial structure image features corresponds to the output of CNN backbone X_i., which is flattened to generate a sequence vector $X_i' \in \mathbb{R}^{L\times d}$, where $L = H \times W$. Then, a 2D sine positional encoding E_{sin_i} is used to encode the position information of generated sequence vector, including 2D bone positional encoding E_{bpos_i}, which are added onto X_i', as shown below:

$$X = [X_i' + E_{sin_i} + E_{bpos_i}] \in \mathbb{R}^{nL\times d}, \tag{2}$$

where n represents the number of views used for the proposed architecture. The standard attention mechanism $\xi(.)$, is fed with an uniform embedding (see Fig. 1), which is built by concatenating two views (i.e., X_1' and X_2') and computed using Eq. 2.

A. Positional Encoding

The used attention mechanisms require the position and order of input sequence. For this, the positional encoding denoted as E_{sin_i} is computed for each individual view following the Eq. 3 proposed by the original transformers in [32]. The relative or absolute position of the obtained features from CNN backbone is used as input to the attention mechanisms. Since that both (obtained features and positional encoding) have the same dimension d_{model}, then both could be summed previously. The following equation shows the positional encoding corresponding to one view:

$$PE_{(pos,2m)} = sin(pos/10000^{2m/d_{model}}). \tag{3}$$

where pos and m correspond to the position information and its index respectively, and d_{model} represents the dimension of feature vector.

B. Bone Position Encoding

In order to improve the attention mechanism on the image, a Bone position encoding denoted as E_{bpos_i} is proposed, which obtain information about position and orientation of body bones, mainly, those bones whose body joints have more probability to be occluded due to the natural human body pose. As first step, a person detector is considered to obtain the bounding box of persons, as shown below:

$$\beta_i = \delta_{detector}(I_i), \tag{4}$$

where the bounding box of detector person is represented as β. The detector person available in the state-of-the-art corresponds to $\delta_{detector}(.)$, I is the current image and i corresponds to the *i-th* view (i.e., view 1 or view 2).

The contour line of human pose from the input image is obtained by using a CNN. For this, the input image is transformed to a gray scale representation, including a set of filters, which are combined to set the weight of CNN. The output of CNN is mapped to other range of values (i.e., between 0 and 255). In order to generate the new image, a new function is implemented, where the plain back background and the cropped contour-image are fused. The obtained coordinates of bounding box of detected person in the input image are used to place cropped contour-image on plain black background in the same image coordinate system (x, y) that original position. The formulate is defined as:

$$\begin{aligned} Ic_i &= \triangle_{crop}(\beta_i(I_i)) \in \mathbb{R}^{h_1,w_1}, \\ NI_i &= \triangle_{image}(\delta_{\square}(.) \in \mathbb{R}^{h,w}, Ic_i \in \mathbb{R}^{h_1,w_1}), \end{aligned} \tag{5}$$

where $\triangle_{crop}(.)$ is a function to crop the contour-image according to the bounding box previously obtained, and h_1 and w_1 are the new height and width of the contour-image after cropping. In order to build a plain black background with original height h and weight w of input image, $\delta_{\square}(.)$ is used. $\triangle_{image}(.)$ is a function that allows to fuse the obtained plain black background and cropped contour-image Ic_i to get a new image (see Fig 2).

A new grid denoted as $G_{h_2\times w_2}$ is built to identify the importance level of each pixel on new image. h_2 and w_2 are set a 32, and correspond to number of row and column of the grid. The intensity value of pixel in the new image near to 255 are considered as important since they could be part of the contour lines of the body pose, and if the pixels are near to 0 then they are considered as irrelevant. As a result of this process, a vector denoted as $V_{(h_2\times w_2),d}$ where d corresponds to the information of each row and column position of each pixel evaluated in the new image, and its intensity value respective. This vector is used as input to the neural network to learn to map these features to ground truth of relevant pixel that make up the bones of body joints more complex, which is formulated as:

$$\begin{aligned} V_{ecd_i} &= \lambda_{\square}(G_{h_2\times w_2}(NI_i)) \in \mathbb{R}^{(h_2\times w_2),d}, \\ \varphi_i &= MLP(V_{ecd_i}) \in \mathbb{R}^{(h_2\times w_2),d}, \end{aligned} \tag{6}$$

where $\lambda_{\square}(.)$ is a function that allows to identify the importance of each pixel in the new image previously obtained in Eq. 5, considering the usage of the new matrix $G_{h_2\times w_2}$ detailed above, and whose output is denoted as V_{ecd_i}. The output of the new image of the *i-th* view is denoted as φ_i, which corresponds to the bone position encoding after applying a multi-layer perceptron to the information obtained in V_{ecd_i} (see Fig 3).

C. Head

The head is built by using the output of Attention Module, which is denoted as $\tilde{X} = \xi(X) \in \mathbb{R}^{(h_2\times w_2),d}$, to predict *K*

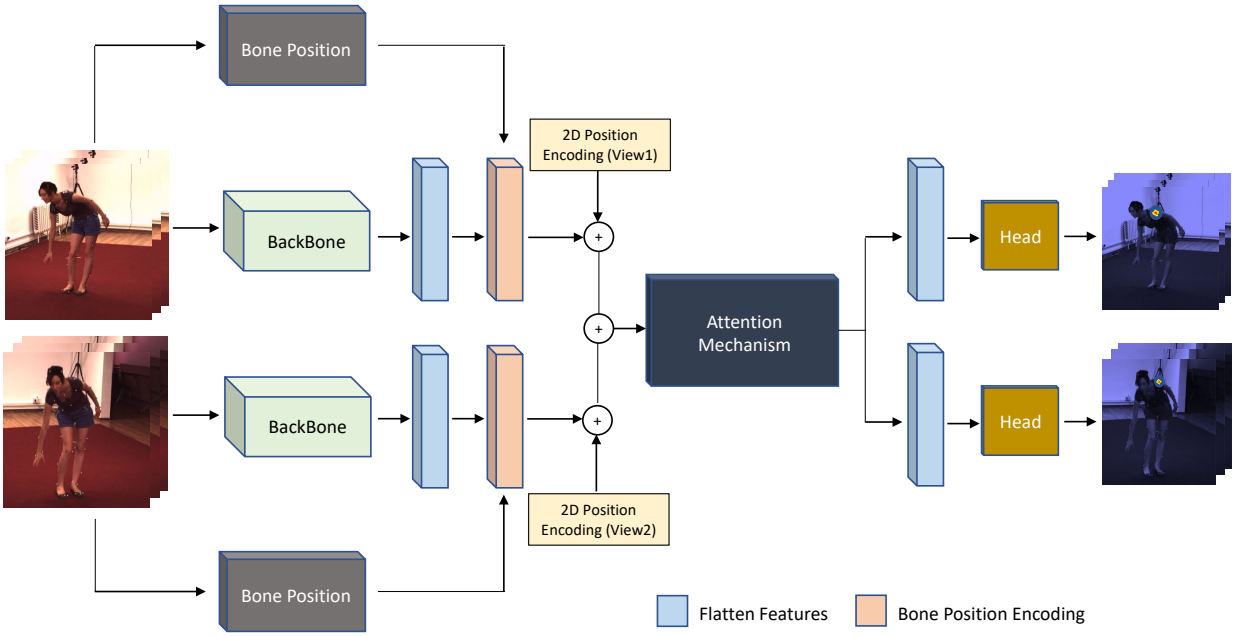

Fig. 1. Overview of proposed architecture using an attention mechanism. Bone position encoding, which helps to guide the architecture to be more precise, and 2D position encoding are used to feed the proposed approach.

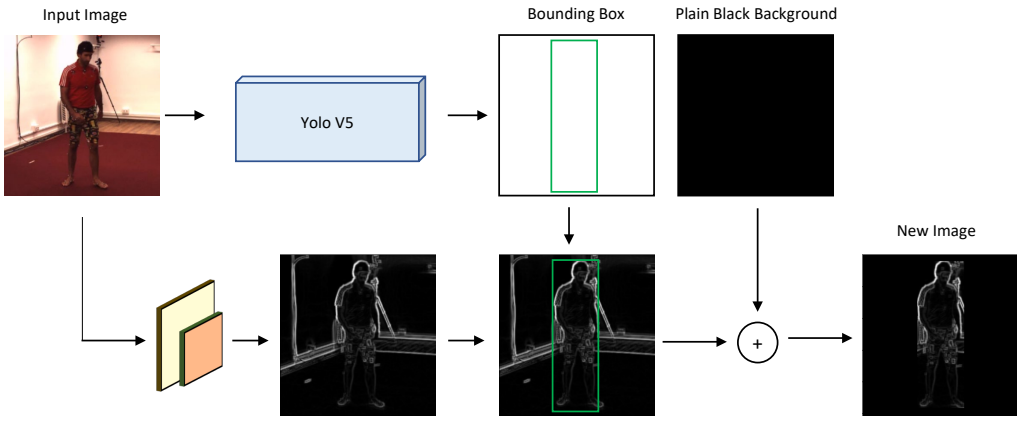

Fig. 2. Overview of general process to get a new image from input image after obtaining the bounding box and contours.

keypoints heatmaps of each view. As a first step, $\tilde{X}$ is split into $\tilde{X}_1$ and $\tilde{X}_2$, and finally, a reshaping is performed to $\tilde{X}_i \in \mathbb{R}^{K\times H^*\times W^*}$, where i represents each view available for the architecture, and H^* and W^* correspond to $H_i/4$ and $W_i/4$ respectively.

In order to reduce the channel dimension of $\tilde{X} = \xi(X)$ from d to K, the equation 7 is applied. Additionally, one deconvolution and 1×1 convolution layer are used. An additional bilinear interpolation or a 4×4 transposed convolution is considered when H_i and W_i are not equals. The equation is formulated as:

$$\Omega_i = H_{\square}(\perp(\tilde{X})) \in \mathbb{R}^{K\times H^*\times W^*}, \tag{7}$$

where $\perp(.)$ corresponds to split the output of the attention module for each view, and $H_{\square}(.)$ is a function that allows to get the heatmaps of body joints from them, whose result are saved in Ω_i.

D. Loss Function

During the learning process of relevant pixels between the bone position encoding generated for the images of *i-th* view and its ground truth, a loss function of bone position encoding is used. This function helps to minimize the error, and is defined as

$$Loss_{bpe} = \frac{1}{L \times d} \sum_{i=1}^{N} \|\varphi_i - \hat{p}_i\|_2 , \tag{8}$$

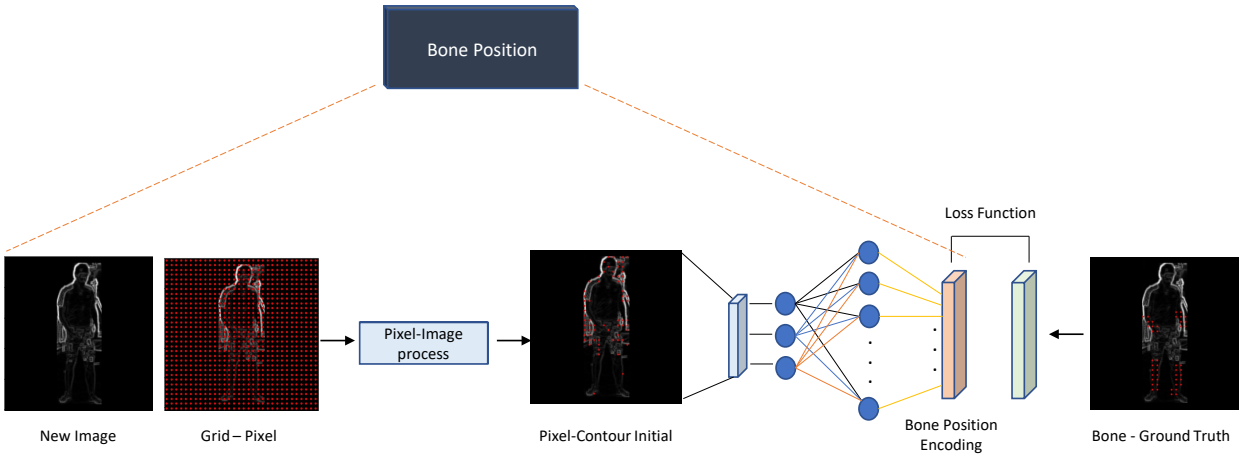

Fig. 3. Details of an attention mechanism where the *bone position encoding* is obtained by using an initial contour of a person, which is used to feed the neural network.

where L represents a matrix $(h_2 \times w_2)$, N corresponds to the number of images, φ_i corresponds to the bone position encoding of images of i-th view obtained from Eq. 6, and $\hat{p}_i$ is the ground-truth of features of the bone in the images of i-th view.

Finally, the Mean Square Error (MSE) loss is applied to obtain the general error of learning process of the proposed architecture. For this, the MSE is computed between the outputs of Head denoted as Ω_i obtained from Eq. 7 and the ground truth heatmap of 2D body joints of input images defined as 2D Gaussian centering around each keyjoint and denoted as $\hat{M}_i \in \mathbb{R}^{K \times H^* \times W^*}$. The proposed architecture is trained end-to-end by using the equation defined below:

$$Loss = \frac{1}{H^* \times W^*} \sum_{i=1}^{N} \left\| \Omega_i - \hat{M}_i \right\|_2 + Loss_{bpe}. \qquad (9)$$

IV. Experiments Results

For the experiments results, one large-scale pose estimation public dataset is used, including metrics to evaluate the performance of the proposed model. This section describes both of them, dataset and used metrics, including the obtained results.

A. Dataset

The used dataset for learning process of proposed architecture is Human3.6m. This dataset is one of the largest publicly available human pose estimation dataset. Four synchronized and calibrated digital cameras are used to capture all scene from different points of view. Different actions performed by the persons are used during the training and testing process of proposed model. The scheme used for these validations are the same as the learning process.

B. Metrics

In order to evaluate the performance of the proposed model, Joint Detection Rate (JDR) metric is used. This metric allows to measure the percentage of $successfully$ detected joints, considering a threshold. The Euclidean distance is also computed to estimated the accuracy of estimated joints in term of distance error.

C. Training

As an initial part of the learning process of proposed architecture, the weights of pretrained Transpose proposed by [33], using MS-COCO dataset [34], are used to initialize the proposed architecture and finetune it on the Human3.6m dataset, since the multi-view pose datasets are quite limited to train it from scratch.

Pytorch is used to implemented the proposed architecture, which is trained with NVIDIA Titan XP GPU and Intel Core I9 3.3GHz CPU. Following the settings in [35], Adam optimizer is used to train the network, including a learning rate of 10^{-3} and decays at 10-th and 15-th epoch with ratio 0.1. A back size of 16 (i.e., eight human poses simultaneously captured from two different points of view) is used for training model. A pre-processing step is performed over the given dataset, where all images are cropped according to the bounding box, which has all the four sides of equal length with the person in the center for keeping the aspect ratio, and then, resizes them to 256×256 pixels. For learning process of bone encoding position of each image input, the relevant pixels of certain body part are estimated to be also used as ground-truth.

A set of 156k images is used to feed to the architecture during the training process, which is trained until 20 epochs; it takes about 144 hours. The pre-processing mentioned above has been also used during the evaluation phase. In the evaluation a set of 8k images has been considered.

D. Results and Comparisons

Quantitative results for human pose estimation using Attention Modules are depicted in Table I. The evaluations of proposed architecture, referred to as Cross view Feature Bone with AM, are compared with two models, Transpose proposed

TABLE I

COMPARISON OF 2D POSE ESTIMATION ACCURACY ON HUMAN3.6M DATASET USING JDR(%) AS METRIC. " − ": THESE ENTRIES WERE ABSENT. ∗ TRAINED AGAIN BY [35]. R50 AND R152 ARE RESNET-50 AND RESNET-152 RESPECTIVELY. SCALE IS THE INPUT RESOLUTION OF THE NETWORK. PARAM CORRESPONDS TO THE NUMBER OF TRAINABLE PARAMETERS OF MODELS. AM MEANS ATTENTION MODULE.

	net	scale	param	shlder	elb	wri	hip	knee	ankle	root	head	Avg
Sum epipolar line [31]	R152	320	-	91.36	91.23	89.63	96.19	94.14	90.38	-	-	-
Max epipolar line [31]	R152	320	-	92.67	92.45	91.57	97.69	95.01	91.88	-	-	-
Transpose with AM [33]	R50	256	5M	95.2	92.2	88.4	98.8	96.9	91	100	99.5	95.25
Cross-View fusion * [35]	R50	320	525M	95.6	95.0	93.7	96.6	95.5	92.8	96.7	96.2	95.26
Cross-View fusion * [35]	R50	256	235M	86.1	86.5	82.4	96.7	91.5	79.0	100	95.5	89.71
Cross view Feature Bone with AM (Ours)	R50	256	5M	95.4	92.2	88.8	98.5	96.7	91	100	99.6	95.28

by [33] and Cross-View fusion proposed by [31], including their different variants by using the JDR metric. Cross-view fusion and their variants (sum and max epipolar line) use the concept of epipolar line for enhancing the performance of models, which considers information of other views to improve the learning process; and thus, obtain better results. The details of models such as network architecture, scale of images and trainable parameters are denoted as "net", "scale" and "param" respectively. The results for each body joints such as shoulder, elbow, wrist, hip, knee, among others, were obtained by using the mean of body joint on both sides of human body, i.e., left and right sides.

Table I shows that Cross view Feature Bone using Attention Module outperforms the previous work on most of body joints, especially those networks and scales of images that are the same than the proposed architecture. If the trained architecture is compared with Cross-View fusion proposed by [31] and retrained by [35] using images of size (256x256 pixels), the improvement is most significant, mainly in body joints such as shoulder, elbow, wrist, knee and ankle. An increment from 86.1% to 95.4%, from 86.5% to 92.2%, from 82.4% to 88.8%, from 91.5% to 96.7% and from 79.0% to 91.0%, respectively, are obtained. Even, when the scale of images in Cross-view fusion model increases to 320x320 pixels, the proposed architecture has a slight advantage. This slight advantage is similar when it is compared with Transpose architecture proposed by [33]. It is important to consider that the body joints with most impact about performance of model are elbow, wrist and ankle, which have major probability of having some type of occlusions due the natural body pose. The bone position encoding as additional information helps to improve these results. On the other hand, large computing power is required when the number of trainable parameters of the proposed architecture increments from 5M to 235M and 525M, which depend of backbone used as network and the scale of images.

Additionally, the accuracy of prediction of each body joint of proposed architecture is obtained by using median euclidean distance error, and they are compared with Transpose proposed by [33]. Note that JDR metric is used to determine if a body joint is considered as successful prediction taking into account a threshold. However, this metric does not give any reference of accuracy of each body joint.

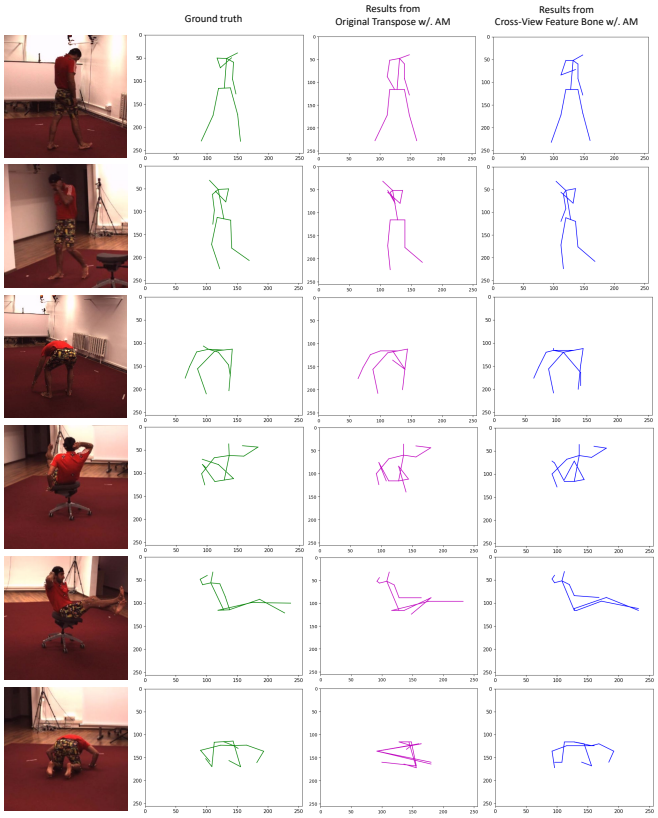

Fig. 4. Challenging poses, Cross view Feature Bone using attention module takes advantage when the feature bone of one view is merged with feature body pose of other view with respect to Transpose architecture proposed by [33], which use a single view. AM means attention module.

The results presented in Table II show that an improvement in accuracy happens in shoulder, wrist, hip, knee. The median Euclidean distance errors for these joints are 1.18%, 2.58%, 0.28% and 1.57% with respect to the results obtained with

TABLE II
COMPARISON OF AVERAGE MEDIAN EUCLIDEAN DISTANCE ERRORS BETWEEN CROSS VIEW FEATURE BONE WITH AM AND TRANSPOSE WITH AM PROPOSED BY [33] ON HUMAN3.6M. AM MEANS ATTENTION MODULE.

Model	shlder	elb	wri	hip	knee	ankle	root	head	Avg
Transpose with AM [33]	5.10	**4.35**	4.65	3.59	4.45	**5.61**	1.24	2.90	3.99
Cross view Feature Bone with AM (Ours)	**5.04**	4.40	**4.53**	**3.58**	**4.38**	5.74	**1.24**	**2.90**	**3.97**

Fig. 5. Other challenging poses where Cross view Feature Bone using attention module takes advantage with respect to the architecture proposed by [33]. AM means attention module.

Transpose architecture proposed by [33]. Fig. 4 5 and present some challenging poses where Cross view Feature Bone architecture takes advantage of information about position and orientation of bones in the image plane with respect to the approach proposed by [33].

A quantitative analysis shows that both proposals present similar body joint accuracy, mainly for the visible joints for body poses such as head and root. However, the accuracy obtained by proposed architecture for body joints that have a high capacity of rotation/mobility such as shoulder, wrist or knee, are better than the prediction obtained from the model proposed by [33].

V. CONCLUSIONS

The proposed approach addresses the challenging problem of the human pose estimation when the joints are occluded. This proposed is motivated by the reduced information of occluded body joints due the natural body pose, mainly, when only one view is available to capture the scene. The attention modules have allowed to design a lightweight architecture, including the integration of relevant features of bone in image plane from both views (i.e., self view and reference view) into learning process. In spite of low number of trainable parameters used for learning process, the proposed architecture has shown appealing results respect to the obtained results of models of state-of-art, mainly, if the numbers of trainable parameters are considered. The manuscript shows how the important features of body joints captured from other views, can be fused to estimate occluded body joints more accurately. It is important to note that the accuracy of body joint estimation is the base to solve others related problems such as surveillance, action recognition, healthcare, among others. Future work will be focused on extending the usage of attention module including the geometry of the scene and extrinsic parameters of the cameras, and thus, improve the human pose estimation.

REFERENCES

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 7291–7299.

[2] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1653–1660.

[3] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[4] C. Tian, Y. Xu, L. Fei, J. Wang, J. Wen, and N. Luo, "Enhanced cnn for image denoising," *Transactions on Intelligence Technology*, vol. 4, no. 1, pp. 17–23, 2019.

[5] M. Wu, H. Yue, J. Wang, Y. Huang, M. Liu, Y. Jiang, C. Ke, and C. Zeng, "Object detection based on rgc mask r-cnn," *Image Processing*, vol. 14, no. 8, pp. 1502–1508, 2020.

[6] J. L. Charco, B. X. Vintimilla, and A. D. Sappa, "Deep learning based camera pose estimation in multi-view environment," in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2018, pp. 224–228.

[7] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 4724–4732.

[8] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *International Conference on Computer Vision*. IEEE, 2017, pp. 2334–2343.

[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 transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 5686–5696.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[12] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021, pp. 1–16.

[13] H. Wang, Y. Zhu, H. Adam, A. Yuille, and L.-C. Chen, "Max-deeplab: End-to-end panoptic segmentation with mask transformers," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 5459–5470.

[14] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 8737–8746.

[15] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 12 294–12 305.

[16] L. Huang, J. Tan, J. Meng, J. Liu, and J. Yuan, "Hot-net: Non-autoregressive transformer for 3d hand-object pose estimation," in *International Conference on Multimedia*. Association for Computing Machinery, 2020, p. 3136–3145.

[17] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *International Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 1954–1963.

[18] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang, "Pix2vox: Context-aware 3d reconstruction from single and multi-view images," in *International Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2690–2698.

[19] H. Sarmadi, R. Muñoz-Salinas, M. Berbís, and R. Medina-Carnicer, "Simultaneous multi-view camera pose estimation and object tracking with squared planar markers," *IEEE Access*, pp. 22 927–22 940, 2019.

[20] J. L. Charco, A. D. Sappa, B. X. Vintimilla, and H. O. Velesaca, "Camera pose estimation in multi-view environments: From virtual scenarios to the real world," *Image and Vision Computing*, vol. 110, p. 104182, 2021.

[21] C. Tang, Y. Ling, X. Yang, W. Jin, and C. Zheng, "Multi-view object detection based on deep learning," *Applied Sciences*, p. 1423, 2018.

[22] X. Luo and F. Li, "Stacked hourglass networks based on polarized self-attention for human pose estimation," in *Second IYSF Academic Symposium on Artificial Intelligence and Computer Engineering*, International Society for Optics and Photonics. SPIE, 2021, pp. 543 – 548.

[23] H. Xia and T. Zhang, "Self-attention network for human pose estimation," *Applied Sciences*, 2021.

[24] X. Wang, J. Tong, and R. Wang, "Attention refined network for human pose estimation," *Neural Processing Letters*, p. 2853–2872, 2021.

[25] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, and Z. Wang, "Tfpose: Direct human pose estimation with transformers," *arXiv preprint arXiv:2103.15320*, pp. 1–15, 2021.

[26] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3d human pose estimation with spatial and temporal transformers," in *International Conference on Computer Vision*. IEEE, 2021, pp. 11 636–11 645.

[27] Z. Wan, Z. Li, M. Tian, J. Liu, S. Yi, and H. Li, "Encoder-decoder with multi-level attention for 3d human shape and pose estimation," in *International Conference on Computer Vision*. IEEE, 2021, pp. 13 013–13 022.

[28] H. Shuai, L. Wu, and Q. Liu, "Adaptively multi-view and temporal fusing transformer for 3d human pose estimation," *ArXiv*, vol. abs/2110.05092, 2021.

[29] T. Wang, J. Zhang, Y. Cai, S. Yan, and J. Feng, "Direct multi-view multi-person 3d pose estimation," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021, pp. 13 153–13 164.

[30] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 466–481.

[31] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross view fusion for 3d human pose estimation," in *International Conference on Computer Vision*. IEEE, 2019, pp. 4342–4351.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[33] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," *International Conference on Computer Vision*, pp. 11 782–11 792, 2021.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[35] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu, "Epipolar transformers," in *Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7779–7788.