

EBA-PRNetCC: An Efficient Bridge Attention-Integration PoseResNet for Coordinate Classification in 2D Human Pose Estimation

Ali Zakir¹^a, Sartaj Ahmed Salman¹^b, Gibran Benitez-Garcia¹^c and Hiroki Takahashi^{1,2}

¹Department of Informatics, Graduate School of Informatics and Engineering,
The University of Electro-Communications, Tokyo, Japan

²Artificial Intelligence Exploration/Meta-Networking Research Center,
The University of Electro-Communications, Tokyo, Japan

Keywords: 2D Human Pose Estimation, EBA-PRNetCC, MLP, EBA, COCO Dataset.

Abstract: In the current era, 2D Human Pose Estimation has emerged as an essential component in advanced Computer Vision tasks, particularly for understanding human behaviors. While challenges such as occlusion and unfavorable lighting conditions persist, the advent of deep learning has significantly strengthened the efficacy of 2D HPE. Yet, traditional 2D heatmap methodologies face quantization errors and demand complex post-processing. Addressing this, we introduce the EBA-PRNetCC model, an innovative coordinate classification approach for 2D HPE, emphasizing improved prediction accuracy and optimized model parameters. Our EBA-PRNetCC model employs a modified ResNet34 framework. A key feature is its head, which includes a dual-layer Multi-Layer Perceptron augmented by the Mish activation function. This design not only improves pose estimation precision but also minimizes model parameters. Integrating the Efficient Bridge Attention Net further enriches feature extraction, granting the model deep contextual insights. By enhancing pixel-level discretization, joint localization accuracy is improved. Comprehensive evaluations on the COCO dataset validate our model's superior accuracy and computational efficiency performance compared to prevailing 2D HPE techniques.

1 INTRODUCTION

Human Pose Estimation (HPE) stands as a significant challenge within the Computer Vision (CV) domain, with its significance emphasized by a multitude of practical applications. Over time, the effort for precise HPE has fostered a deep engagement with Deep Learning (DL) and Convolutional Neural Networks (CNNs) among the CV community. The current state-of-the-art methods have indeed achieved commendable success, delivering impressive qualitative and quantitative results. This progress naturally sparks interest regarding the potential advancements in the coming years and the room for further improvement in this domain. However, a practical disparity exists. Despite the high accuracy achieved by recent models, many applications have yet to reap the benefits of these advancements fully. The essence of this limitation hinges on two factors: (a) the as-

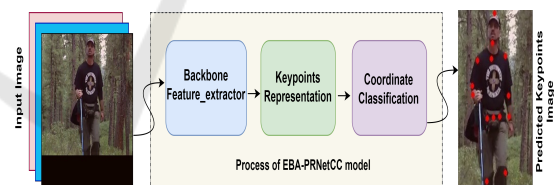





Figure 1: Overall framework of the proposed EBA-PRNetCC model.

sumption of abundant computational resources such as GPUs, memory, and power, which often contradicts the reality for many applications, and (b) the imperative of maintaining accuracy, particularly in critical domains like autonomous driving where there is minimal tolerance for error when transitioning to more compact, memory-efficient methods (Zakir et al., 2023a). The existing literature presents numerous methods demonstrating superior performance on rigorous benchmarks such as MPII (Andriluka et al., 2014), LSP (Johnson and Everingham, 2010), and Common Objects in Context (COCO) (Lin et al., 2014). However, a significant gap persists as none

^a <https://orcid.org/0000-0002-3187-9551>

^b <https://orcid.org/0000-0001-9344-6658>

^c <https://orcid.org/0000-0003-4945-8314>

attain this high accuracy level under memory and computational power constraints. The primary objective of our effort is to bridge this gap, proposing advancements over the current state-of-the-art methods in these challenging settings. Simultaneously, the ongoing expansion and integration of CV in devices like smartphones and surveillance systems have led to a constant flow of image and video data. The information extracted regarding human actions and events is highly valuable. HPE aims to identify and record different joints within the human body, thus estimating a person's posture through the spatial positions of body parts such as arms and head, referred to as keypoints (Zhang et al., 2023).

Over the past decade, the CV field has extensively explored the automatic understanding of HPE. 2D HPE is a foundation for numerous advanced CV tasks, including transitioning from 2D to 3D poses, detecting human actions, and improving Human-Computer Interaction (HCI). The intricacy of 2D HPE arises from challenges such as obscured keypoints, difficult lighting conditions, and the demanding task of real-time deployment due to the model's large number of parameters (Chen et al., 2022). Traditionally, the initial approaches utilized standard techniques such as probabilistic graphical models, which were heavily dependent on manually designed features, thereby limiting the model's adaptability and effectiveness. The emergence of DL addressed these constraints by enabling automatic feature extraction from data. Specifically, the advancements made by CNNs in 2D HPE have inspired a multitude of deep learning techniques (Cao et al., 2017).

Recently, methods predicated on 2D heatmaps have emerged as the predominant approach but often falter due to quantization errors, leading to issues like poor performance at low resolutions, high computational demands, the need for multiple upsampling layers, and complex post-processing steps such as non-maximum suppression and heatmap smoothing. These factors significantly contribute to these methods' high computational demand and complex post-processing (Xiao et al., 2018; Yang et al., 2021; Salman et al., 2023b). Contrary to existing solutions, our research introduces a coordinated classification approach for 2D HPE, presenting an alternative to the conventional 2D heatmap-based method. This paper's primary objective is to improve prediction accuracy and optimize the model by reducing the number of parameters addressing the computational constraints. This initiative guides the focus toward enhancing the precision of existing models without exacerbating the computational demands, providing a balanced approach to tackling the challenges in 2D

HPE.

We propose the EBA-PRNetCC model, as depicted in Figure 1. This model utilizes an efficient version of ResNet as its foundational backbone for primary feature extraction (He et al., 2016). In our adaptation, we have preserved only the convolutional structures, omitting both the average pooling and the final fully connected layers. The selection of ResNet34 is intentional, aiming for a balance between computational efficiency and model complexity, especially when compared against more complex variants like ResNet50, 101, and 152. Recognizing the trade-off between precision and model size, our approach employs specific strategies to counter potential reductions in accuracy. We have integrated the EBANet—a sophisticated version of Bridge Attention (BA) (Zhao et al., 2022)—into ResNet34. This integration emphasizes BA's role in the architecture, enhancing communication between layers. By serving as a channel for feature transference and regulating the prominence of specific features, it refines the network's focus on critical patterns in the feature representation, efficiently addressing potential information bottlenecks. One fundamental refinement resides in the model's head: introducing a dual-layer Multi-Layer Perceptron (MLP) complemented by the Mish activation function. This enhancement not only pares down the parameter volume but also boosts pose estimation accuracy, a feat we attribute to the non-linear properties of the MLP and the Mish function's enriched gradient dynamics. Upon capturing the keypoint representations using the backbone, EBA-PRNetCC processes the vertical and horizontal coordinates individually, leading to its final predictions. Our suggested model partitions every pixel into several bins, reducing quantization inaccuracies and offering precision that surpasses single-pixel localization.

The threefold contribution of the proposed model can be summarized as follows:

- We proposed EBA-PRNetCC as a transition from the conventional 2D heatmap-based methods. This approach adopts a coordinated classification strategy, utilizing an efficient ResNet34 for foundational feature extraction. This ensures a notable balance between keypoint prediction accuracy and computational efficiency, as demonstrated by a controlled parameter count.
- We integrated the EBANet into ResNet34, enhancing the model's capability in feature recognition and contextual understanding, which is crucial for human pose estimation. An essential modification is the redesigned head, comprised of a two-layer MLP with Mish activation. This change reduces parameters and improves accu-

racy. Our pixel segmentation approach also minimizes quantization errors, optimizing joint localization precision.

- A comprehensive evaluation on the COCO dataset affirmed the effectiveness of our proposed method. Our approach demonstrated superior precision when assessing the results and required fewer computational resources than current 2D human pose estimation methods.

The structure of the paper is laid out as follows for ease of understanding: Section 2 provides an overview of prior studies in this field. Section 3 describes the techniques and principles behind our EBA-PRNetCC. In Section 4, we discuss the experimental setup and the specific details of our implementation. Section 5 and 6 offers an in-depth analysis of our results. Concluding the paper, Section 7 summarizes the discussions and points out potential paths for future studies.

2 RELATED WORK

In 2D HPE, DL methods have gained importance due to their expertise in extracting features ranging from simple to sophisticated. Initially, 2D HPE research primarily centered on regression-based approaches (Tian et al., 2019; Nie et al., 2019). These models aimed to locate keypoint coordinates directly, but their inconsistent reliability hindered widespread adoption. The landscape evolved significantly when (Li et al., 2021) introduced the Residual Log-Likelihood (RLE). This method leveraged the benefits of normalizing flows and delivered performance comparable to premier heatmap techniques. Following this, much of the HPE research shifted towards utilizing two-dimensional Gaussian distribution heatmaps for precise joint coordinate mapping (Cao et al., 2017; Cai et al., 2020; Cheng et al., 2020). Pioneering contributions from researchers like (Tompson et al., 2014) and (Newell et al., 2016) were contributory to this shift, leading to innovative architectures such as the celebrated hourglass design. As investigations in the field deepened, a strong push emerged for methods that maintained high-resolution outputs from start to finish in the computation process (Sun et al., 2019). However, even with these advances, heatmap-based strategies faced difficulties. Notably, quantization error remained a persistent challenge, becoming even more evident in scenarios with lower resolutions.

A significant challenge in 2D heatmap-based HPE is quantization errors, particularly prominent in smaller-dimensional heatmaps. Recognizing this,

many researchers have developed innovative solutions (Zakir et al., 2023b; Salman et al., 2023a; Salman et al., 2023c). (Zhang et al., 2020) introduced a post-processing method using Taylor expansion to improve heatmap distribution approximation. In a distinctive approach, (Yin et al., 2020) proposed transitioning from traditional 2D heatmaps to a more compact 1D heatmap format, incorporating special adjustable layers. They further enhanced the resolution of these 1D heatmaps using additional deconvolution layers. While many HPE research efforts focus on such techniques, some avant-garde methods stand out. For instance, in facial landmark detection, the use of 1D heatmap techniques is gaining traction, as seen in pioneering works like that of (Yin et al., 2020), who introduced a pivotal 1D heatmap regression method, setting a new benchmark in the field.

(Chen et al., 2021) proposed Pix2Seq, an innovative technique bridging the field of object detection and linguistic modeling. SimCC (Li et al., 2022) explored a departure from conventional heatmap-centric strategies for HPE in a related development. Their approach, compatible with both CNN and Transformer-based HPE systems, eliminates the need for a distinct Transformer decoder during predictions (Mao et al., 2021). In contrast, our work, anchored on EBA-PRNetCC, presents a refined solution to classical heatmap-based HPE approaches, bypassing the demands of resource-heavy post-processing.

3 EBA-PRNetCC

In the field of 2D HPE, given an RGB image or a video frame labeled as I , the goal is to identify the pose: the pose \mathbf{P} of any individual is represented in this visual content. This posture, expressed as \mathbf{P} , is characterized by a set of N specific keypoints. Each keypoint is denoted by a two-dimensional coordinate (x_n, y_n) . The number of keypoints, N , can vary based on the dataset used for training a model. Thus, our objective is to pinpoint the pose $\mathbf{P} = \{\mathbf{P}_i\}_{i=1}^N$ for every k individual within the input, as described in Algorithm 1. proposed

Our EBA-PRNetCC model, as illustrated in Figure 2, clearly demonstrates our proposed approach. As depicted, we have deliberately selected ResNet34 for an optimal balance between computational demands and model complexity. The illustration emphasizes the incorporation of Efficient Bridge Attention (EBA) Net within the ResNet34 structure, emphasizing its role in enhancing inter-layer communication. The diagram also captures the model's unique methodology in processing vertical and horizontal co-

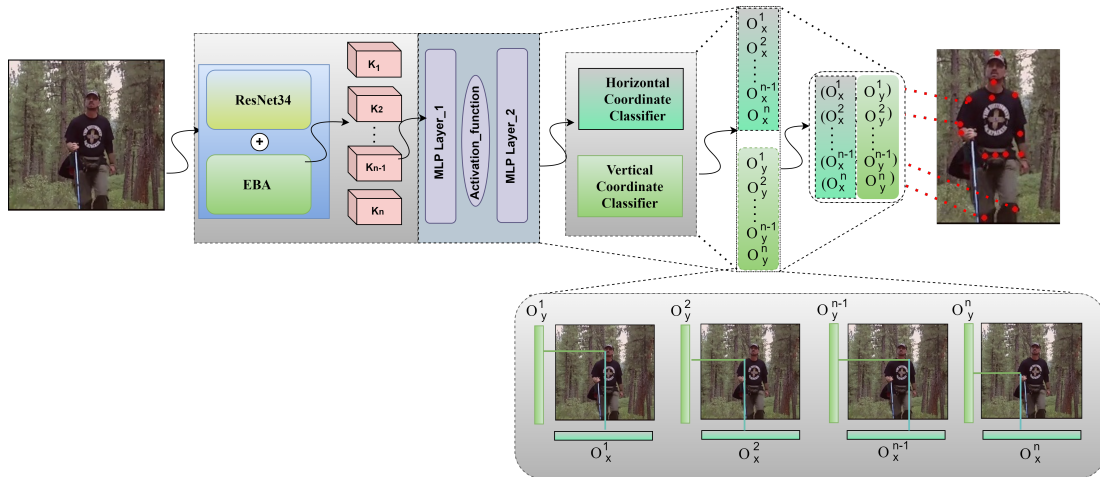


Figure 2: Detail architecture of our proposed EBA-PRNetCC model for coordinate classification in 2D HPE.

Data: RGB image or video frame I

Result: Posture P of every individual in I

Initialize $P = \emptyset$ (Set to store postures of individuals);

Detect number of individuals k in image I ;

for $i = 1$ to k **do**

 Detect K keypoints for individual i in I ;

 For each keypoint n , obtain its coordinates (x_n, y_n) ;

 Store the keypoints for individual i as P_i ;

 Add P_i to P ;

end

Return P ;

Algorithm 1: 2D Human Pose Estimation (HPE).

ordinates, concluding in its final predictive outcomes. For a more comprehensive understanding of its intricate design and functionality, we discuss each of its components in greater detail in the subsequent subsections.

3.1 Backbone Enhancement Using Adapted ResNet

Autoencoder network architectures are gaining attraction within research communities, especially in tasks requiring complicated annotations or detailed dataset labeling. These architectures demonstrate a distinctive capability to manage and optimize feature representations effectively. In response to these requirements, we adopted an autoencoder design that logically reduces the resolution of feature representations in stages. This step-by-step reduction optimizes computational efficiency and retains broad spatial details, guaranteeing the preservation of essential information. One of the primary advantages of this design

lies in its capability to both intensify and recover the original spatial clarity of the feature maps, making the processing more robust and accurate. In contrast, architectures like the Hourglass frequently produce feature maps with more constrained dimensions than their original inputs. Challenges can emerge when resizing these condensed outputs to their initial dimensions. One of the significant issues is the risk of quantization, where the understated details might get overshadowed or lost. Further complicating the process are potential biases during data transformation, which can introduce errors. This is particularly noticeable in scenarios where models misunderstand data due to operations like horizontal mirroring, emphasizing the importance of precise and bias-free data handling.

In our search to address the abovementioned challenges, we gravitated toward integrating ResNet34 into our proposed model. This decision was grounded in the model's inherent efficiency, particularly highlighted when considering the parsimonious parameter count of ResNet34 compared to its denser counterparts, such as ResNet 50, 101, and 152. Our revision of ResNet (He et al., 2016) allowed us to omit the average pooling and fully connected stages, setting the stage for a more efficient processing approach. Instead of these, our model features four distinct ResNet blocks as shown in Figure 3. Each block is a sequence of convolution operations, batch normalization, advanced Mish activation, and max pooling operations. The input process through our model begins with a convolutional layer promptly followed by a pooling layer, effectively reducing the spatial dimensions of the feature maps by a factor of 2. These refined feature maps then serve as inputs to the subsequent ResNet blocks. Complexities within these blocks involve convolutional layers, further reducing feature dimensions by a factor of two, although the initial

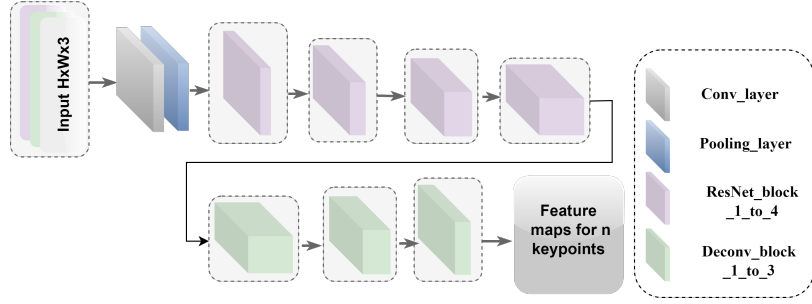


Figure 3: Efficient ResNet-Based architecture with integrated deconvolutional modules for precise feature representation for n Keypoints.

ResNet block stands as an exception to this behavior. We incorporated three deconvolutional modules, each enhanced by batch normalization and employing the Mish activation function. A notable feature of these deconvolutional stages is their ability to upscale the resolution of the feature maps iteratively. This process continues until the feature maps align with the spatial dimensions of the original input, guaranteeing a thorough and detailed feature representation.

The EBA-PRNetCC model is tailored to process images of dimensions $H \times W \times 3$. In this format, H and W symbolize the image’s height and width, respectively. The third dimension, represented by 3, corresponds to the familiar RGB color scheme, breaking down into Red, Green, and Blue channels. ResNet34 plays a pivotal role within this setup as the primary mechanism for extracting detailed features from the image. As the image progresses through the analytical depth of the ResNet34 backbone, the model adeptly identifies and maps out n specific keypoint representations, each corresponding to its designated keypoint within the image.

3.2 Efficient Bridge Attention Net

In deep convolutional networks, attention mechanisms serve as essential enhancements, enabling models to prioritize salient features in data dynamically. The BA methodology is particularly distinctive among these mechanisms. It offers solutions to challenges inherent in conventional attention techniques, primarily by ensuring an optimized utilization of information across neural networks. A notable attribute of the BA approach is its adept feature compression within the attention layer. A critical insight underlying BA-Net is the recognition in the observation that while advanced layers in neural structures primarily interpret high-level features, preliminary layers provide contextual features. Due to their layer-specific focus and computational constraints, traditional attention methodologies often by-

pass the nuanced information available in these initial layers. Addressing this gap, BA-Net introduces a strategic ‘bridge,’ facilitating a comprehensive integration of features throughout the network’s depth. Expanding on the foundational concepts of our approach, we will probe further into the detailed operations of the EBA module as illustrated in Figure 4. Given a block, an output from a function is characterized as $\mathbf{X}_i \in \mathcal{R}^{C_i \times H \times W}$. Here, C_i denotes the feature map’s depth at the i^{th} layer. An essential operation in our methodology is the application of Global Average Pooling (GAP) to these outputs, reducing their spatial dimensions to $C_i \times 1 \times 1$. This condensed representation is crucial for facilitating subsequent feature integration and processing. Post-GAP, these features undergo transformation through matrices sized $C_i \times \frac{C_n}{r}$, producing “squeezed” features. Here, C_n is the n number of channels in the feature map at the deepest or current layer under consideration. This particular dimensional transformation is essential to guarantee the smooth integration of features from diverse layers. However, this procedure can induce variations in feature distributions. Batch normalization BN is employed to counteract this, standardizing these distributions and bolstering their non-linear representations. Mathematically, the integration process is expressed as:

$$\mathbf{S}_i = BN_i(\mathbf{W}_i(\text{GAP}(F_i))) \quad (1)$$

F_i is the feature map at the i -th layer of the network. The transformation matrix is denoted by W_i , where S_i encapsulates the squeezed feature. For the BA mechanism, the comprehensive feature representation, denoted as $I_{BA}(\cdot)$, emerges as a summation of these features:

$$I_{BA}(\cdot) = \sum_{i=n-q}^n S_i \quad (2)$$

where features from the $(n-q)$ -th layer up to the n -th layer are summed together. The value of q determines how many previous layers’ features are included in this summation. This composite feature then advances to the generation phase G , resulting in

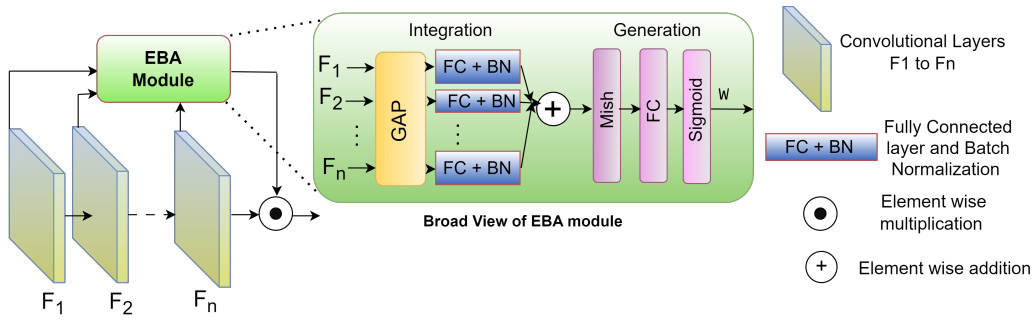


Figure 4: Visualization of the EBA module's feature integration and weights W generation mechanisms.

the final attention weights W :

$$W = G(I_{BA}) \quad (3)$$

However, simply fusing the features isn't the final step. The generation sequence in the module, distinguished by the use of the Mish activation function, further refines this fused feature. This sequence culminates with a sigmoid activation, ensuring the resultant attention weights are bounded between 0 and 1. These weights are instrumental in modulating the original feature maps, emphasizing salient regions while deemphasizing others. The EBA module feature processing is describe in Algorithm 2.

Its integration within the ResNet-34 architecture deserves special mention. ResNet34, with its series of convolutional layers, offers a fertile ground for EBA-Net's capabilities. Both the BasicBlock and Bottleneck classes within ResNet34 incorporate the BA-Net. During their forward passes, features are extracted at various stages, processed, and then modulated by the EBA-Net derived attention weights. This integration ensures that as data flows through the ResNet34 architecture, the model continually refines its focus, leveraging the combined wisdom of current and preceding features. The result is a more attentive, context-aware ResNet-34, poised to discern intricate patterns with heightened precision.

3.3 Head and Coordinate Classification

As visualized in Figure 2, our improved model incorporates a refined head classification technique. One essential refinement resides in the model's head: introducing a dual-layer MLP complemented by the Mish activation function. This enhancement not only pares down the parameter volume but also boosts pose estimation accuracy, a feat we attribute to the non-linear properties of the MLP and the Mish function's enriched gradient dynamics. Instead of simply appending horizontal and vertical classifiers with a single linear layer for each, we have embedded these MLPs. Each MLP is structured to first transform the

Data: Block output $X_i \in \mathbb{R}^{C_i \times H \times W}$

Result: Attention weights W that modulate the original feature maps

Initialize C_n as the number of channels in the feature map at the deepest/current layer;

Initialize r as the reduction ratio;

/* GAP Function */ Apply Global Average Pooling (GAP) to feature maps X_i to get the condensed representation $C_i \times 1 \times 1$;

/* Integration Function */ For each layer i from $n - q$ to n :

- Apply GAP to X_i to get the condensed representation;
- Transform the features using matrices of size $C_i \times \frac{C_n}{r}$;
- Apply Batch Normalization (BN) to the transformed features;
- Define S_i as $S_i = \text{BN}_i(\mathbf{W}_i \times \text{GAP}(X_i))$;

Initialize I_{BA} to zero;

for $i = n - q$ **to** n **do**
 $S_i = \text{Integration of } X_i$;
 $I_{BA} += S_i$;

end

/* Generation Function */ Apply Mish activation function to I_{BA} ;

Apply sigmoid activation function to the output of Mish;

Define W as the output of sigmoid activation;

Return the resultant attention weights W ;

Algorithm 2: EBA Module Feature Processing.

representation into an intermediate dimension before outputting the classification results. With respect to the ResNet backbone, the resultant keypoint representations are reformatted from $(n, H', W') \rightarrow (n, H' \times W')$ for the classification phase. This modified approach still retains the lightweight nature of the original SimCC head but is designed for enhanced feature capture, especially crucial for intricate object scenarios.

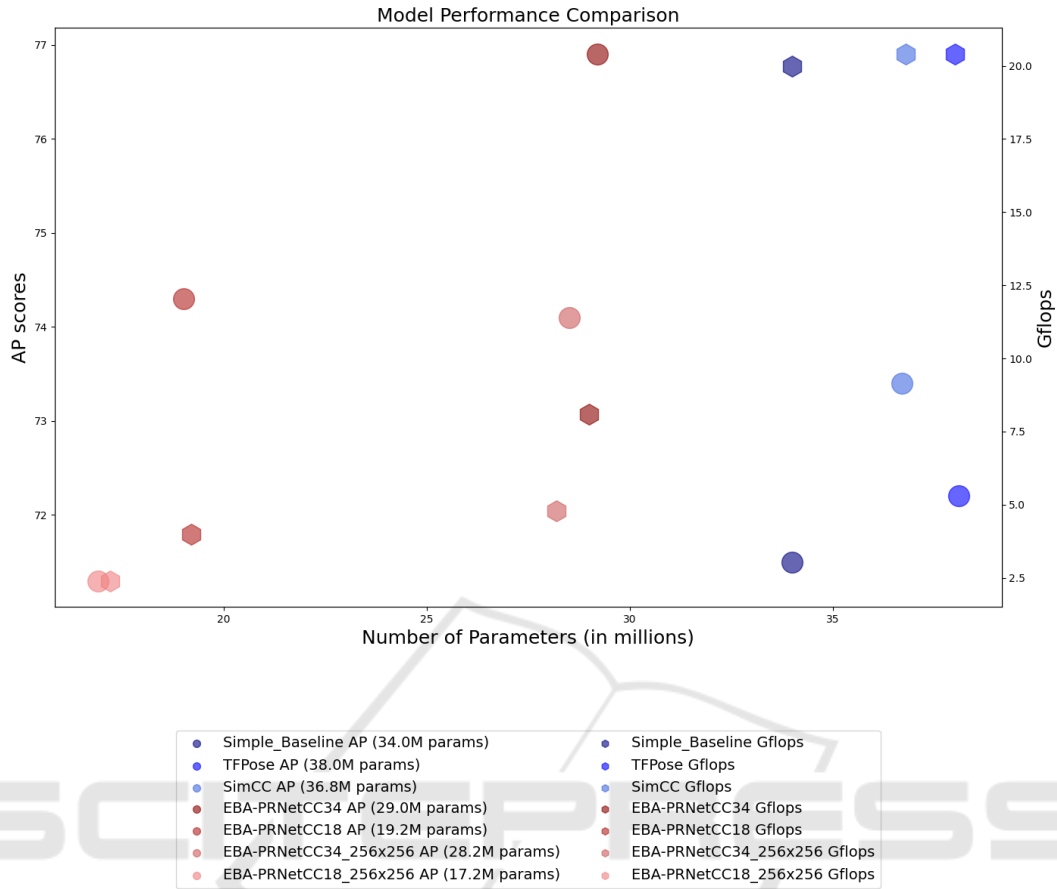


Figure 5: Visual analysis of 2D HPE models in terms of accuracy, parameters, and computational efficiency.

For the classification task, our method is rooted in the principle of evenly converting each continuous coordinate into an integer, which then acts as a training class label: c_x falls within $[1, N_x]$, and c_y within $[1, N_y]$. Here $N_x = W \times k$ and $N_y = H \times k$ specify the bin counts for the horizontal and vertical dimensions, respectively. The division factor, k , is sensibly chosen to be ≥ 1 to curtail quantization discrepancies, assuring granular localization accuracy. In generating the concluding predictions, EBA-PRNet independently carries out vertical and horizontal coordinate classifications, drawing on the n keypoint representations discerned by the backbone. For a given i^{th} keypoint representation, the associated predictions o_x^i and o_y^i are procured through the respective horizontal and vertical classifiers.

3.4 Loss Function

Our 2D HPE EBA-PRNet methodology adopts a distinctive loss function termed KLDiscrctLoss. This loss function utilizes the Kullback-Leibler Diver-

gence (KLD) to measure the closeness between the predicted pose coordinate distributions and the actual ground truth distributions. At its core, this function is an adaptation of PyTorch’s KLDivLoss, emphasizing non-aggregated loss outputs. Before computing the divergence, predictions undergo a transformation via the LogSoftmax layer, rendering them into a log-probability format conducive to the KLD evaluation. During the forward pass, the loss is computed for each joint’s x and y coordinates. For every joint, the predicted coordinates are juxtaposed with the ground truth, and a specific weight modifies the resulting loss. The aggregate loss is then derived by averaging these individual joint losses, providing a holistic assessment of the model’s efficacy in pinpointing human pose coordinates. Mathematically, the loss for the i^{th} joint is expressed as:

$$\text{Loss}[i] = W[i] \times \left(\text{KLD}(\text{logSoftmax}(o_x^i), \text{gt}(o_x^i)) + \text{KLD}(\text{logSoftmax}(o_y^i), \text{gt}(o_y^i)) \right) \quad (4)$$

The overall `KLDiscretLoss` is then computed as the mean of losses across all joints:

$$\text{KLD}_{\text{iscretLoss}} = \frac{1}{N} \sum_{i=1}^N \text{Loss}[i] \quad (5)$$

Here, N represents the total number of joints. $(o_x^i$ and (o_y^i) indicate the predicted x and y coordinates for the i^{th} joint, respectively, while $\text{gt}(o_x^i)$ and $\text{gt}(o_y^i)$ are the corresponding ground truth coordinates. W symbolizes the weight designated for the i^{th} joint. The term $\text{KLD}(A, B)$ defines the KLD between distributions A and B .

4 EXPERIMENTAL SETUP

4.1 Dataset

The COCO keypoint dataset (Lin et al., 2014) occupies a prominent position in the field of 2D HPE. Valued for its extensive range and variety, it has over 200K images. These images, captured across diverse real-world settings, annotate a notable 250K individuals across 17 distinct human joint categories. This variety not only facilitates the training of complex models but also presents challenges due to diverse lighting, occlusions, and complex human poses. For our research, we carefully limited our training to the COCO 2017 training subset, which includes 57K images annotated for 150K individuals. It's essential to highlight that our training routine remained exclusively within this dataset, preserving the integrity and uniformity of our outcomes without the addition of any outside data. Additionally, the dataset allocates 5K images for validation, a vital process in optimizing and enhancing model accuracy. Another 20K images are reserved for testing efforts, facilitating a thorough evaluation of the model's performance in unfamiliar situations.

4.2 Evaluation Metric

We utilize evaluation metrics derived from the Object Keypoint Similarity (OKS) methodology to assess the accuracy of our keypoint localization. OKS offers a measure of the difference between the predicted keypoints and their true ground-truth positions. The mathematical representation of OKS is:

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (6)$$

In this formula, d_i represents the Euclidean distance between the predicted and ground-truth keypoints.

The term s refers to the scale of the person and k_i is a constant specific to each keypoint, accounting for its inherent variability. The variable v_i is an indicator of keypoint visibility, taking a value of 1 if the keypoint is visible and 0 if it's either not visible or not labeled. The function δ acts as an indicator, yielding 1 if its condition is satisfied and 0 otherwise.

Our primary evaluation metric is the Average Precision (AP), which we carefully compute across ten distinct OKS thresholds. For an in-depth analysis, we offer a range of metrics. The AP metric provides a wide-ranging perspective, averaging AP scores over OKS values from 0.50 to 0.95, incremented by 0.05. We further clarify performance at specific OKS values with AP50 and AP75 metrics. We introduce metrics AP(M) and AP(L) for medium and large objects to address different object sizes. Lastly, the Averaging Recall metric, AR values over the selected OKS thresholds, gives a comprehensive insight into the model's recall proficiency.

4.3 Implementation Details

We integrated specific data augmentation techniques to enhance the proposed model's adaptability to spatial rotations and scale variations. These included random horizontal reflections, rotations varying from -30 to +30 degrees, and scale changes within the 0.7 to 1.3 range. We constructed our model using the PyTorch framework. During the training phase, we set the learning rate at 1e-05, established a batch size of 16 for training and testing, utilized six dedicated workers for parallel tasks, and conducted training over 140 epochs. The Mish activation function was our preference over the traditional ReLU, primarily because Mish effectively navigates the challenges of vanishing gradients, a limitation often associated with ReLU. By retaining gradients during operations, Mish promotes both efficient and steady training, especially crucial for deeper network structures, positioning it as an indispensable tool for our model's superior functionality. For optimization purposes, we included the AdamW optimizer, an advanced iteration of the classic Adam method. A defining trait of AdamW is its ability to handle weight decay and learning rate modifications distinctly, facilitating enhanced tuning accuracy and substantially diminishing the tendency for overfitting.

5 RESULTS AND DISCUSSION

In our study, we compared the performance of various pose estimation methodologies, including our

Table 1: EBA-PRNetCC Performance Evaluation in Comparison with Previous Pose Estimation Methods.

Method	Repr.	Backbone	Input	AP	AP(50)	AP(75)	AP(M)	AP(L)	AR
Simple Baseline	Heatmap-based	ResNet-50	384x288	71.5	91.1	78.7	67.8	78.0	76.9
TFPose	Reg.-based	ResNet-50T.		72.2	90.9	80.1	69.1	78.8	-
SimCC	Coord.-based	ResNet-50		73.4	89.2	80.0	69.7	80.6	78.8
EBA-PRNetCC	Coord.-based	ResNet-18		74.3	92.5	81.5	71.5	79.0	77.2
		ResNet-34		76.9	93.5	83.6	73.9	81.6	79.6



Figure 6: Qualitative Results of EBA-PRNetCC for 2D HPE on COCO Dataset Under Viewpoint Changes, Occlusions, and Adverse Imaging Conditions.

Table 2: Comparison of EBA-PRNetCC vs. Previous Methods: Input, AP, Parameters, and Computational Efficiency (GFLOPS).

Method	Input	AP	P(M)	GF
Simple Baseline	384×288	71.5	34.0	20.0
TFPose		72.2	38.0	20.4
SimCC		73.4	36.8	20.4
EBA-PRNetCC18		74.3	19.0	4.0
EBA-PRNetCC34		76.9	29.0	8.1
EBA-PRNetCC18	256x256	71.3	17.0	2.4
EBA-PRNetCC34		74.1	28.0	4.8

proposed model EBA-PRNetCC. The results are presented in Table 1, which summarizes the performance metrics across different models. From Table 1, it is obvious that the Simple Baseline method (Xiao et al., 2018), based on a heatmap representation with a ResNet-50 backbone, achieved an AP of 71.5, AP(50) of 91.1, and AP(75) of 78.7. When considering medium-sized objects AP(M), the method attained 67.8 and 78.0 for larger objects AP(L). The average recall (AR) for this model stands at 76.9. The TFPose (Mao et al., 2021), a regression-based approach with a combined ResNet-50 and Transformer

backbone, slightly outperformed the Simple Baseline with an AP of 72.2, AP(50) of 90.9, and AP(75) of 80.1. However, the AR for this model is not provided. SimCC (Li et al., 2022), another competitive model, uses a coordinate-based representation with a ResNet-50 backbone. It showed a promising AP of 73.4 and an AR of 78.8. Notably, its performance on larger objects, AP(L), is 80.6, which is marginally better than the aforementioned models. Our proposed method, EBA-PRNetCC, was tested with two different backbones: ResNet18 and ResNet34. With the ResNet18 backbone, EBA-PRNetCC achieved an AP of 74.3, which is higher than both the Simple Baseline and TFPose methods.

The performance further improved with the ResNet34 backbone, reaching an AP of 76.9, making it the best-performing model among the ones tested. Furthermore, the EBA-PRNetCC with ResNet34 achieved impressive results on AP(50) and AP(75) with scores of 93.5 and 83.6, respectively. The AR for this variant is 79.6, which is also the highest among the models presented. Table 2 offers a deeper dive, focusing on the balance between AP, model pa-

Table 3: Performance comparison of EBA-PRNetCC with SLP and MLP Head.

Method	Head	Input	AP	AP(50)	AP(75)	AP(M)	AP(L)	AR	Param(M)	GFLOPS
EBA-PRNetCC	SLP	384×288	76.3	93.5	82.6	73.2	80.9	79.0	35	8.2
	MLP		76.9	93.5	83.6	73.9	81.6	79.6	29	8.1

Table 4: Performance comparison of EBA-PRNetCC variants with different backbones.

Method	Backbone	Input	AP	AP(50)	AP(75)	AP(M)	AP(L)	AR
EBA-PRNetCC	ResNet18	256×256	71.3	91.4	78.1	68.2	75.8	74.5
	ResNet34		74.3	92.5	81.5	71.5	79.0	77.2

rameters, and computational complexity in Gflops. The Simple Baseline (Xiao et al., 2018), with its 384x288 input size, demands approximately 34.0M parameters and 20.0 Gflops. TFpose (Mao et al., 2021), despite having a slightly better AP of 72.2, requires more parameters (38.0M) while maintaining a computational complexity of 20.4 Gflops, similar to SimCC’s requirements (Li et al., 2022). Our EBA-PRNetCC models, notably EBA-PRNetCC34 and EBA-PRNetCC18, stand out for their efficiency. EBA-PRNetCC34, with its ResNet34 backbone, not only tops in AP with 76.9 but does so with just 29M parameters, consuming only 8.1 GFLOPS. When powered by ResNet-18, EBA-PRNetCC18 retains a competitive AP of 74.3, but with a drastic reduction in parameters to 19M and computational need to 4.0 GFLOPS. Even with a reduced input size of 256x256, both models continue to impress. Their performance, combined with reduced parameters and GFLOPS, underlines their efficiency. our EBA-PRNetCC models, both with ResNet34 and ResNet18 backbones, offer a compelling balance between accuracy and efficiency, we visualized the AP, GPLOPS, and number of Parameters of our models and other previous models in Figure 5.

Figure 6 provides a thorough qualitative evaluation of our EBA-PRNetCC for 2D HPE, illustrating its effectiveness on the COCO dataset under a variety of challenging conditions. The model adeptly manages complex situations, including alterations in viewpoint, diverse occlusions, blurry imagery, extreme lighting variations, and complex human-object interactions. The displayed outcomes emphasize the model’s ability to precisely comprehend and adapt to the nuanced dynamics of the human kinematic chain, ensuring reliable and accurate pose estimation throughout these demanding scenarios.

6 ABLATION STUDY

To understand the influence of the enhancements made to the model’s head, we conducted an ablation study comparing two key configurations of our EBA-

PRNetCC model: one employing a Single Layer Perceptron (SLP) and the other utilizing a dual-layer MLP with the Mish activation function. The results of this study are presented in Table 3. EBA-PRNetCC-SLP used a ResNet34 backbone, this version with an appended SLP achieved an AP of 76.3 and required 35M parameters with 8.2 GFLOPS computational complexity. While EBA-PRNetCC-MLP introducing a dual-layer MLP with Mish activation, this variant not only improved the AP to 76.9 but also reduced the parameters to 29M while maintaining a similar computational complexity of 8.1 GFLOPS. The results highlight the efficiency and performance gains achieved by embedding an MLP with Mish activation in the model’s head, making EBA-PRNetCC more effective without increasing computational demands.

Table 4 presents the performance of the EBA-PRNetCC model with a reduced input size of 256×256, using both ResNet-18 and ResNet34 as backbones. With the ResNet18 backbone, the model achieved an AP of 71.3. Its AP(50) score was particularly impressive at 91.4, indicating its ability to accurately detect more obvious poses. The AR for this configuration stood at 74.5. Switching to the ResNet-34 backbone brought about significant improvements. The AP increased to 74.3, and the AP(50) rose to an impressive 92.5. The overall recall also saw an enhancement, registering at 77.2. These results highlight the model’s consistent performance, even when the input size is changed.

7 CONCLUSIONS AND FUTURE WORK

This paper introduces EBA-PRNetCC, a novel methodology that shifts from the conventional 2D heatmap-centric techniques to a sophisticated, coordinated classification strategy in 2D HPE. Leveraging a modified ResNet34 architecture, our system achieves reliable keypoint detection while reducing the number of parameters. A crucial enhancement in our model’s design lies in its head incorporating a dual-

layer MLP accentuated by the Mish activation function. This deliberate inclusion reduces the parameters and amplifies pose estimation accuracy—a result we link to the non-linear attributes of the MLP and the gradient-rich dynamics of the Mish function. Moreover, the EBA infusion within ResNet34 enhances our model’s feature extraction capabilities, granting it deeper contextual insights. By emphasizing pixel-level discretization, we curtail quantization irregularities and boost joint localization precision. Experimental results produce EBA-PRNetCC superior performance on the COCO dataset, attributed to its refined feature mapping, optimal activation function, and sophisticated optimization techniques. In the future, we aim to adopt increasingly efficient architectures and expand training over varied datasets to enhance model generalization.

REFERENCES

- Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.
- Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., Zhang, X., Zhou, X., Zhou, E., and Sun, J. (2020). Learning delicate local representations for multi-person pose estimation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, pages 455–472. Springer.
- Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Real-time multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.
- Chen, H., Feng, R., Wu, S., Xu, H., Zhou, F., and Liu, Z. (2022). 2d human pose estimation: A survey. *Multi-media Systems*, pages 1–24.
- Chen, T., Saxena, S., Li, L., Fleet, D. J., and Hinton, G. (2021). Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*.
- Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., and Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Johnson, S. and Everingham, M. (2010). Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*.
- Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., and Lu, C. (2021). Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11 025–11 034.
- Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., and Xia, S.-T. (2022). Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pages 89–106. Springer.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*.
- Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., and Wang, Z. (2021). Tfpose: Direct human pose estimation with transformers. *arXiv preprint arXiv:2103.15320*.
- Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII*, pages 483–499. Springer.
- Nie, X., Feng, J., Zhang, J., and Yan, S. (2019). Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6951–6960.
- Salman, S. A., Zakir, A., Benitez-Garcia, G., and Takahashi, H. (2023a). Acenet: Attention-driven contextual features-enhanced lightweight efficientnet for 2d hand pose estimation. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE.
- Salman, S. A., Zakir, A., and Takahashi, H. (2023b). Cascaded deep graphical convolutional neural network for 2d hand pose estimation. In *International Workshop on Advanced Imaging Technology (IWAIT) 2023*, volume 12592, pages 227–232. SPIE.
- Salman, S. A., Zakir, A., and Takahashi, H. (2023c). Sdfposegraphnet: Spatial deep feature pose graph network for 2d hand pose estimation. *Sensors*, 23(22):9088.
- Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703.
- Tian, Z., Chen, H., and Shen, C. (2019). Directpose: Direct end-to-end multi-person pose estimation. *arXiv preprint arXiv:1911.07451*.
- Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27.
- Xiao, B., Wu, H., and Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481.
- Yang, S., Quan, Z., Nie, M., and Yang, W. (2021). Transpose: Keypoint localization via transformer. In *Pro-*

- ceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11802–11812.
- Yin, S., Wang, S., Chen, X., Chen, E., and Liang, C. (2020). Attentive one-dimensional heatmap regression for facial landmark detection and tracking. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 538–546.
- Zakir, A., Salman, S. A., Benitez-Garcia, G., and Takahashi, H. (2023a). Aeca-prnetcc: Adaptive efficient channel attention-based poseresnet for coordinate classification in 2d human pose. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE.
- Zakir, A., Salman, S. A., and Takahashi, H. (2023b). Sahf-lightposeresnet: Spatially-aware attention-based hierarchical features enabled lightweight poseresnet for 2d human pose estimation. In *International Conference on Parallel and Distributed Computing: Applications and Technologies*, pages 43–54. Springer.
- Zhang, F., Zhu, X., Dai, H., Ye, M., and Zhu, C. (2020). Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102.
- Zhang, T., Lian, J., Wen, J., and Chen, C. P. (2023). Multi-person pose estimation in the wild: Using adversarial method to train a top-down pose estimation network. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.
- Zhao, Y., Chen, J., Zhang, Z., and Zhang, R. (2022). Banet: Bridge attention for deep convolutional neural networks. In *European Conference on Computer Vision*, pages 297–312. Springer.