Motion-Aware Heatmap Regression for Human Pose Estimation in Videos

Inpyo Song[1] , **Jongmin Lee**[1] , **Moonwook Ryu**[2] and **Jangwon Lee**[1]
[1]Department of Immersive Media Engineering, Sungkyunkwan University, Republic of Korea
[2]Electronics and Telecommunications Research Institute, Republic of Korea
{songinpyo, eel8, leejang}@skku.edu, moonwook@etri.re.kr

Abstract

We present an approach to solving 2D human pose estimation in videos. The problem of human pose estimation in videos differs from estimating human poses in static images since videos contain a lot of motion related information. Thus, we investigate how to utilize by the information of the human body movements across in a sequence of video frames for estimating human poses in videos. To do this, we introduce a novel heatmap regression method what we call motion-aware heatmap regression. Our approach computes motion vectors in joint keypoints from adjacent frames. We then design a new style of heatmap that we call ***Motion-Aware Heatmaps*** to reflect the motion uncertainty of each joint point. Unlike traditional heatmaps, our motion-aware heatmaps not only consider the current joint locations but also account how joints move over time. Furthermore, we introduce a simple yet effective framework designed to incorporate motion information into heatmap regression. We evaluate our motion-aware heatmap regression on PoseTrack(2018, 21) and Sub-JHMDB datasets. Our results validate that the proposed motion-aware heatmaps significantly improve the precision of human pose estimation in videos, particularly in challenging scenarios such as videos like sports game footage with substantial human motions. (Code and related materials are available at https://github.com/Songinpyo/MTPose.)

1 Introduction

In this paper, we address the problem of 2D human pose estimation in videos, which aims to detect and localize the major joints in the human body (e.g., elbows, wrists, etc.) from a sequence of video frames. It is essential for building a wide range of intelligent systems such as video surveillance systems and autonomous driving systems [Chen *et al.*, 2020; Munea *et al.*, 2020; Human-Machine-Interfaces-Trend-Report,]. However, despite the importance of the problem, much of the research up to now has more focused on a single-frame based human pose estimation method [Newell *et al.*, 2016; Fang *et al.*, 2017; Xiao *et al.*, 2018; Sun *et al.*, 2019; Cheng *et al.*, 2020; Luo *et al.*, 2021; Xu *et al.*, 2022b], and comparatively little attention has been paid to human pose estimation in videos. This is probably because single-frame based models can be directly applied to videos without any modification, and they seem to show reasonable performance. Such approaches, however, have missed an opportunity to improve their performance by taking advantage of temporal information across video frames.

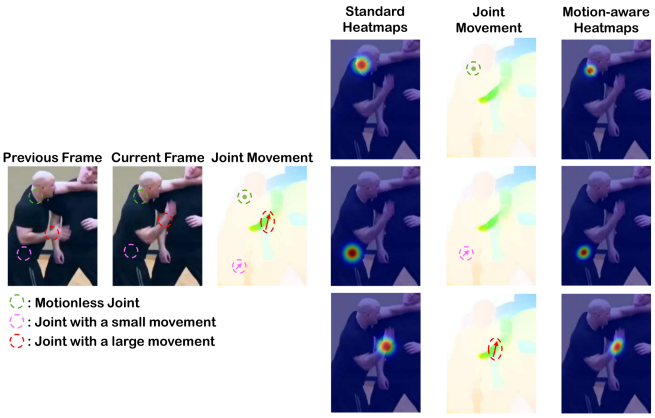

Figure 1: We propose a new style of heatmaps what we call ***Motion-Aware Heatmaps*** to reflect the motion uncertainty of each joint keypoint for estimating human poses in videos. Our motion-aware heatmaps not only consider the current joint locations but also account for the temporal dynamics of joint movements.

Therefore, more recently, questions have been raised about the use of the single-frame based methods in videos for human pose estimation. The main disadvantage of the single-frame based techniques is that their performance rapidly drops when videos include large motions since they cause motion blur and pose occlusions. Thus, to date, researchers have investigated multi-frame based approaches to estimate human body configuration in videos [Bertasius *et al.*, 2019; Liu *et al.*, 2021; Liu *et al.*, 2022; Jin *et al.*, 2022; Feng *et al.*, 2023b]. This previous research attempted to utilize additional temporal information from neighboring video frames for estimating human poses in the current video frame. One line of this research proposes to employ Recurrent Neural Networks (RNNs)-based models like Long short-term memory (LSTM), GRU (Gated Recurrent Units)

or 3D Convolutions to obtain better human motion representations by aggregating the spatio-temporal features from adjacent frames in a video sequence [Luo *et al.*, 2018; Wang *et al.*, 2020]. Another line of work suggests to explicitly combine the joint keypoints of neighboring video frames based on their pose heatmaps as post-processing [Liu *et al.*, 2021; Liu *et al.*, 2022; Jin *et al.*, 2022; Feng *et al.*, 2023b; Feng *et al.*, 2023a].

However, previous multi-frame based studies do not take into account to use the joint motion cue, the movements of the human joint keypoints in human motions, that can help improve the performance of video-based human pose estimation. We believe that the joint motion cues can provide strong evidence to detect and localize the joint keypoints, especially, in videos with substantial motions. Therefore, in this research we explore the ways to predict the joint configuration of the human body in videos by leveraging this useful information in the joint motion cues. To capture the joint motion cues, we propose a novel heatmap regression approach using a new style of heatmaps that we call ***Motion-Aware Heatmaps*** and motion data intergrating architecture.

To be specific, we adjust standard deviation of the Gaussian kernels for each joint keypoint based on the magnitude of a joint movement vector that represents the scales of the each joint keypoint movements of human body. In addition, we adjust the shape of the gaussian kernels according to the direction of the motion vector of each joint keypoint. The motivation behind this new style heatmap is that we observed the previous approaches often failed to estimate human poses in videos when they have fast-moving subjects. Especially, in those type of videos, fast moving joints of the subjects are likely to be hard to estimate its locations in images rather than their motionless joints. Thus, we attempt to represent those different difficulties (uncertainties) by adjusting the scales of the standard deviation in our new style of heatmap. Furthermore, we adjust the shape of the Gaussian kernels to represent the direction of each joint movement since it is likely to move in the same direction within a short period of time. Finally, we employ a simple yet effective motion intergrating architecture **MTPose** (**M**o**T**ion-aware **Pose** regression) to estimate the human posture in current frame by leverage motion information from predicted motion-aware heatmaps using our model. This study contributes to the growing area of multi-frame based human pose estimation by introducing a novel heatmap regression method to the existing human pose estimation research in the community. We believe our motion-aware heatmap regression can be used to build various intelligent systems which require to robust estimate postures of people in videos, even they include dynamic human motion and pose occlusions.

The specific contributions of the paper are summarized as follows:

- This is the first study that addresses the uncertainties of human joint keypoints related to its movements on the problem of 2D human pose estimation in videos.
- We propose a new style of heatmaps that reflect the motion uncertainty of each joint point, we then demonstrate that our motion-aware based heatmap regression method can robustly detect joint keypoints with large movements.
- The proposed method is general and straightforward, so that further research could easily adopt our motion-aware heatmaps to take advantage of the useful information from joint motion cues to improve their performance.

2 Related Work

2.1 Single-Frame Based Human Pose Estimation

In the last few years, there has been great progress in estimating 2D human poses in static images [Newell *et al.*, 2016; Xiao *et al.*, 2018; Sun *et al.*, 2019; Cheng *et al.*, 2020; Luo *et al.*, 2021; Li *et al.*, 2021; Xu *et al.*, 2022a; Xu *et al.*, 2022b]. Prior to the work of Tompson [Tompson *et al.*, 2014], previous studies on this single-frame based human pose estimation have been carried out to directly regress the coordinates of joint keypoints in the images [Toshev and Szegedy, 2014; Tompson *et al.*, 2015]. However, these early methods suffer from some serious shortcomings since they do not take account of uncertainty of each joint and the models are hard to be generalized (these early models are often overfitted during training). Thus, much of the recent research on human pose estimation use the indirect method called heatmap regression to infer the probability of the existence of joint keypoints at each pixel in the images. These researches are divide into two strategies, *bottom-up* and *top-down*. The top-down approach first employs a person detector to detect individuals and then estimates the pose for each bounding box of person independently. Conversely, the bottom-up approach detects individual body parts and associates these parts with the persons identified in the image.

Since these heatmap regression-based models require the ground-truth heatmaps for training, researchers usually generate the ground-truth heatmaps by putting 2D Gaussian kernels with the same standard deviation on all joint keypoints in images based on the ground-truth annotations. However, [Luo *et al.*, 2021] proposed a new approach what they called scale-adaptive heatmap regression to adaptively adjust the standard deviation based on human scales on the images in a bottom-up. But, scale-adaptive heatmap regression might not help improve the performance of top-down models since they use fixed size of person on the image frame, making the adjustment of standard deviation according to human scale is less impactful.

Distinctly we focus on the joint motion cue, the movements of the human joint keypoints in human motions, to adjust of the standard deviation instead of human scales on the images which may not affect for top-down approaches. In addition, we attempt to adjust the shape of the gaussian kernels according to the direction of the joint motion cue whereas others do not. Furthermore, our approach can be applied to the both top-down and bottom-up human pose estimation approaches in the same way since we adjust the standard deviation and rotation based on the magnitude and direction of the motion vector at the joint position.

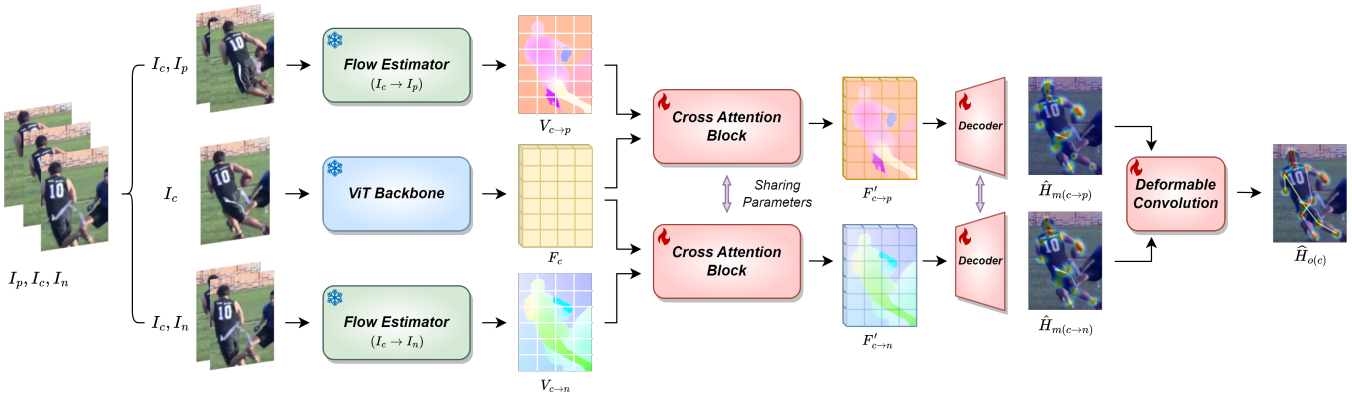

Figure 2: The architecture of MTPose, a simple and effective framework integrating motion information into heatmap regression.

2.2 Multi-Frame Based Human Pose Estimation

Single-Frame based methods show unsatisfactory performance in estimating human poses for video inputs with heavy motion blur and occlusions. Consequently, the development of multi-frame has become essential to utilize the temporal information across adjacent video frames. [Liu *et al.*, 2021; Dang *et al.*, 2022; Liu *et al.*, 2022; Jin *et al.*, 2022; Feng *et al.*, 2023b; Feng *et al.*, 2023a]. Additionally, propagation-based methods have also been proposed to sequentially process the prediction results [Xu *et al.*, 2021; Nie *et al.*, 2019; Luo *et al.*, 2018].

Much of this line of research concentrate on developing explicit algorithms to integrate temporal information from adjacent frames as a post-processing to estimate the human poses in the current frame. For instance, Liu *et al.* attempted to merge three single-frame based pose estimation results obtained from their backbone model, HRNet-w48 [Sun *et al.*, 2019], to get the refined human poses for the current input video frame by computing temporal distances between frames [Liu *et al.*, 2021]. Similarly, [Feng *et al.*, 2023b; Jin *et al.*, 2022] proposed a Transformer-based architecture to leverage temporal context information, utilizing the initially estimated pose heatmaps from the backbone network. [Bertasius *et al.*, 2019; Liu *et al.*, 2022; Feng *et al.*, 2023a] utilize motion information extracted from adjacent frames in the process of pose estimation for temporal information.This approach demonstrates the significance of motion features in pose estimation. However, this approach increases the complexity of the overall model by requiring the design of a separate module for motion extraction and the implementation of a loss function to train this module. In this paper, our network directly produce distinct heatmaps considering joint motion cues in videos. It's applicable without complex motion extraction modules or specialized loss functions

3 Proposed Method

We introduce a novel method, **M**o**T**ion-aware **Pose** regression (**MTPose**), designed for the detection and localization of human body keypoints in videos. Our focus lies in capturing joint motion cues to reflect uncertainty of each joint keypoint according to human motions in the innovative form of heatmaps, referred to as ***Motion-Aware Heatmaps***. We then train our model using these motion-aware heatmaps to leverage all useful information in the joint motion cues represented for estimating human poses in videos.

Method Overview Given a sequence of input videos, we first employ a person detector to identify all human bounding boxes in each frame. Subsequently, each bounding box is expanded by 25% around the detected person to ensure consistent tracking within the scene. These bounding boxes serve as inputs for the network. Following this, our MTPose processes three consecutive frames, denoted as I_p, I_c, I_n, corresponding to the previous, current, and next frames, respectively, per person (bounding boxes) extracted from the input videos. The current frame I_c is then input into a Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020] to obtain the image feature F_c, and two pairs of adjacent frames, namely $I_p\&I_c$ and $I_c\&I_n$, are fed into an off-the-shelf optical flow model [Jiang *et al.*, 2021] to obtain optical flow vectors, denoted as $V_{c\rightarrow p}$ and $V_{c\rightarrow n}$. Following this, the extracted features F_c, $V_{c\rightarrow p}$, and $V_{c\rightarrow n}$ are integrated using a cross-attention module followed by a pose decoder to produce motion-aware heatmaps. Finally, these heatmaps are merged via a deformable convolution layer to form the final heatmap.

To train this network, we use three ground-truth heatmaps: two motion-aware heatmaps for past and future frame movements, and one conventional heatmap focusing on the current frame's spatial joint locations. In the next section, we explain how to generate these ground-truth motion-aware heatmaps.

3.1 Motion-aware Heatmaps

To date, the majority of previous human pose estimation studies have utilized 2D Gaussian kernel that has the same standard deviation (σ) for all keypoints in the both x and y axes, leading to circular-shaped heatmaps to train their deep learning models by heatmap regression. These heatmaps, referred to as original heatmaps (H_o) in this paper, are generated using a base standard deviation, σ_0. In contrast, we design novel motion-aware heatmaps (H_m) to represent the uncertainties of human joint keypoints related to its movements. The basic idea of generating the motion-aware heatmaps is illustrated in Figure 1. Our method modifies standard deviations based on the joint movement vectors. So, we begins generation of

ground-truth motion-aware heatmaps by computing the joint movement vectors JM between adjacent frames. The vector $JM \in \mathbb{R}^2$ is derived from the x and y coordinates of joint keypoints, and is calculated as:

$$JM_{c\to p} = J_c^k - J_p^k, \quad JM_{c\to n} = J_c^k - J_n^k \tag{1}$$

Here, J_p, J_c, and J_n denote the human joint keypoint corresponding to the previous, current, and next frames, respectively, and k represents its keypoint index. After that we establish a motion threshold δ to categorize the joints as either motionless or in motion.

Motionless Joints ($|JM| \leq \delta$): If the joint keypoint's position remains relatively stable across consecutive observed image frames, predicting the location of the joint in the next frame becomes more straightforward compared to joints with high movement. Consequently, to capture this characteristic in our motion-aware heatmaps, we generate a circular Gaussian heatmap with a standard deviation of

$$\sigma_x = \sigma_y = \sigma_0 + \frac{|JM| - \delta}{\delta} \tag{2}$$

This reduced standard deviation signifies a narrowed probability range for joint existence due to its small movement.

Joints in Motion ($|JM| > \delta$): For joints undergoing significant motion, our method generates elliptical Gaussian heatmaps. These heatmaps are designed to reflect both the magnitude and direction of each joint's movement, as illustrated in Figure 3. We first dynamically adjust the standard deviations σ_x and σ_y, based on the joint's movement:

$$\sigma_x = \sigma_0 + \frac{|JM| - \delta}{\delta}, \quad \sigma_y = \sigma_0 - \frac{|JM| - \delta}{\delta} \tag{3}$$

resulting in an elliptical heatmap shape that correlates with the joint's motion. The motion-aware heatmaps are then formulated using the following 2D Gaussian kernel:

$$G_{k,i,j} = (e^{-(i-x_k)^2/2\sigma_x^2})(e^{-(j-y_k)^2/2\sigma_y^2})$$
$$s.t. \|i - x_k\|_1 \leq 3\sigma_x, \|j - y_k\|_1 \leq 3\sigma_y$$

Here, $G_{k,i,j}$ denotes the 2D Gaussian kernel to generate motion-aware heatmaps to cover k^{th} keypoint and i and j indicate the coordinate of the keypoint in image frame.

After that the orientation of our heatmap is determined by the joint's movement direction. The rotation angle θ_{JM} is calculated with this formula:

$$\theta_{JM} = \text{atan2}(JM_y, JM_x) \tag{4}$$

which ensures the major axis of the heatmap corresponds with the joint's motion direction. This alignment is crucial for capturing the joint's presence accurately along its motion path while reducing presence probability perpendicular to this path. Finally, we rotate the generated Gaussian Kernels according to the angle θ_{JM}. By implementing this methodology, our heatmaps effectively represent not only the spatial location but also the directional flow of joint movements, providing a holistic view of human motion dynamics. As a result, two motion-aware heatmaps ($H_{m(c\to p)}, H_{m(c\to n)}$) and one original heatmap ($H_{o(c)}$) are generated as ground-truth for our model.

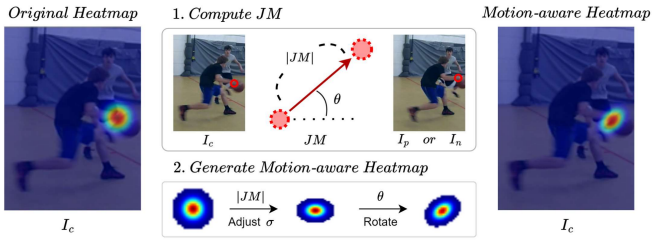

Figure 3: This figure illustrates the generation of ground-truth motion-aware heatmaps for moving joints by adjusting the heatmap shape based on calculated joint movement vectors.

3.2 Motion-aware Features

A motion estimator is essential for incorporating motion information into our framework to enhance its awareness of human motions in videos. In this work, we employ an off-the-shelf optical flow model [Jiang *et al.*, 2021] as our motion estimator to acquire optical flow vectors. The optical flow motion vectors $V_{c\to p}, V_{c\to n} \in \mathbb{R}^{2\times H/d\times W/d}$ are computed from I_c paired with I_p and I_n respectively, utilizing the optical flow model. Here, d represents the downsampling ratio of the patch embedding layer in our ViT backbone. These vectors are then integrated into the feature F_c extracted from the current input frame I_c using the ViT backbone. This integration is achieved through a Multi-Head Cross-Attention (MHCA) layer and a Feed-Forward Network (FFN), after unifying their dimensions via a convolution layer with a kernel size of 1×1. The outcome of this process is a motion-aware feature representation F':

$$F' = F_c + FFN(MHCA(F_c, V, V)) \tag{5}$$

3.3 Heatmap Regression

With the motion-aware feature representation F', our MTPose predicts two motion-aware heatmaps $\widehat{H}_{m(c\to p)}, \widehat{H}_{m(c\to n)}$ using two deconvolution blocks. Each block comprises a deconvolution layer, followed by batch normalization and ReLU activation, effectively upscaling the feature maps by a factor of two. The formulation for this process can be expressed as:

$$\widehat{H}_m = Conv_{1\times1}(Deconv(Deconv(F'))) \tag{6}$$

After obtaining these motion-aware heatmaps $\widehat{H}_{m(c\to p)}, \widehat{H}_{m(c\to n)}$, MTPose generates the final output heatmap $\widehat{H}_c$ using deformable convolutions with five different dilations [Zhu *et al.*, 2019]. Specifically, two motion-aware heatmaps are concatenated to extract offsets $\mathbb{O}$ and masks $\mathbb{M}$ for the deformable convolutions.

$$\begin{aligned}\mathbb{O}_d &= \text{Conv}_{3\times3}(\text{Concat}[\widehat{H}_{m(c\to p)}, \widehat{H}_{m(c\to n)}]) \\ \mathbb{M}_d &= \text{Conv}_{3\times3}(\text{Concat}[\widehat{H}_{m(c\to p)}, \widehat{H}_{m(c\to n)}])\end{aligned} \tag{7}$$

where d corresponds to each of the five dilation levels.

Simultaneously, the average of the two motion-aware heatmaps, $\widehat{H}_{m(avg)}$, is computed to encapsulate both spatial and temporal information. This $\widehat{H}_{m(avg)}$ is then serves as the

input for the Deformable Convolution (DCNv2) process. Applying DCNv2 across various dilations, the convolution results are aggregated to form $\widehat{H}_{o(c)}$, the final heatmap. This aggregation is achieved through a weighted summation:

$$\widehat{H}_{o(c)} = \frac{1}{5} \sum_{d \in 3,6,9,12,15} \text{DCNv2}(\widehat{H}_{m(avg)}, \mathbb{O}_d, \mathbb{M}_d) \quad (8)$$

By utilizing $\widehat{H}_{m(avg)}$, representing the overlapping area of the two motion-aware heatmaps, along with the offset and mask obtained using the two motion-aware heatmaps for DCN, the final heatmaps can be regressed, incorporating motion information from the continuous frame.

3.4 Loss Function

To train our network, we utilize a standard heatmap estimation loss, which is formulated to minimize the Euclidean distance between the predicted heatmaps generated by our model and the corresponding ground-truth for each joint. The loss function is formulated as follows:

$$L(H, \widehat{H}) = \frac{1}{N} \sum_{i=1}^{N} v_i \times ||H^i - \widehat{H}^i||^2$$

$$L_{total} = L(H_{m(c \to p)}, \widehat{H}_{m(c \to p)}) + L(H_{m(c \to n)}, \widehat{H}_{m(c \to n)}) + \gamma * L(H_{o(c)}, \widehat{H}_{o(c)}) \quad (9)$$

In this equation, N represents the total number of joints, v the visibility of each joint, i the index of the joint, and γ the weighting factor of the loss.

It is noteworthy that we can instruct the model to recognize the movement of keypoints using a standard loss function and training procedure, eliminating the need for complex loss functions or models. This is made possible by leveraging the proposed motion-aware heatmaps. This introduces a novel aspect to comprehend joint movements in the learning process, utilizing the familiar heatmap format widely employed in prior approaches. Consequently, this enhances the model's ability to accurately predict human poses in videos, even in the presence of substantial human motions.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conducted three sets of experiments on the three different public benchmark datasets to evaluate the proposed motion-aware heatmap regression approach.

PoseTrack For our evaluation, we utilized two versions of the PoseTrack dataset: PoseTrack2018 [Iqbal *et al.*, 2017] and PoseTrack21 [Doering *et al.*, 2022]. These datasets are key benchmarks in multi-person pose estimation and tracking within video contexts. PoseTrack2018 comprises 1,138 videos, enriched with 153,615 pose annotations. It offers a comprehensive platform for assessing the performance of pose estimation models in diverse scenarios. PoseTrack21 is a extension of previous dataset, this version expands the dataset by including annotations for smaller figures and individuals in crowded settings, totaling 177,164 pose annotations. Both PoseTrack datasets were originally developed to address two core tasks: multi-person pose estimation (Task 1) and multi-person pose tracking (Task 2). Our research focuses exclusively on Task 1, using the Average Precision (AP) metric for evaluation purposes.

Sub-JHMDB The Sub-JHMDB dataset [Jhuang *et al.*, 2013] is a subset of the larger JHMDB collection. It encompasses 316 videos, each averaging 35 frames, across 12 action classes, all with accompanying pose annotations. For assessing 2D pose estimation performance, we apply the Percentage of Correct Keypoints (PCK) metric. This metric calculates the proportion of predicted keypoints that are within a certain threshold distance from their actual locations. We have adopted thresholds of 0.2, 0.1, and 0.05, as per [Zeng *et al.*, 2022; Jin *et al.*, 2023], to cover a range from soft to strict evaluation standards.

4.2 Implementation Details

In developing our MTPose model, we tailored key parameters for optimal performance. The model operates on images resized to 256×192 image size. We used the ViT Backbone initialized from pretrained by [Xu *et al.*, 2022b] and finetuned on PoseTrack2018, PoseTrack21 and Sub-JHMDB datasets. For the frame interval, we set it to 2 for PoseTrack2018 and Sub-JHMDB, while PoseTrack21 uses an interval of 1 due to its high density of small person poses and the associated challenges in capturing joint movements over longer intervals. We have set the motion threshold δ and the default standard deviation σ_0 for Gaussian heatmaps at 3 for all datasets. The training setup involves a loss weigth γ of 3/4, a batch size of 64, a learning rate of 3e-4 using the AdamW optimizer, and is conducted on a single NVIDIA Tesla V100 GPU.

4.3 Comparison With SOTA Methods

We evaluated our MTPose against leading video-based human pose estimation methods on PoseTrack validation sets (AP metric) and Sub-JHMDB dataset (PCK metric).

Results on PoseTrack2018 We first compared our MTPose with existing state-of-the-art methods on the PoseTrack2018 validation set. The evaluation results are presented in Table 1. The table highlights that MTPose achieved a notable advancement, demonstrating an mAP of 89.0. Our model outperformed the prior best-performing method, OTPose [Jin *et al.*, 2022], by a substantial margin of 4.8 mAP. Particularly noteworthy improvements were observed in challenging joints, with gains of 7.1 mAP for the wrist and 8.3 mAP for the ankle. MTPose also achieves an operational efficiency of approximately 14 frames per second on the PoseTrack18 dataset.

We also conducted an additional experiment to assess the effectiveness of our model concerning joint movement scales. This analysis involved computing the average magnitudes of each joint movement within the dataset (where we also normalized them, considering variations in person size using the size of the bounding box). By doing this, we were able to sort each joint according to the magnitude of its movements. Figure 4 shows the results of this evaluation. Our MTPose also exhibited significantly higher gains in joints with larger average movements, highlighting its proficiency in capturing dynamic joint motion.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
AlphaPose [Fang *et al.*, 2017]	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
MDPN [Guo *et al.*, 2018]	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
Dynamic-GCN [Yang *et al.*, 2021]	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
PoseWarper [Bertasius *et al.*, 2019]	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
PT-CPN++ [Yu *et al.*, 2018]	82.4	88.8	86.2	79.4	72.0	80.6	76.2	80.9
DCPose [Liu *et al.*, 2021]	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
DetTrack [Wang *et al.*, 2020]	84.9	87.4	84.8	79.2	77.6	79.7	75.3	81.5
FAMI-Pose [Liu *et al.*, 2022]	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
DiffPose [Feng *et al.*, 2023b]	85.0	87.7	84.3	81.5	81.4	82.9	77.6	83.0
TDMI-ST [Feng *et al.*, 2023a]	86.7	88.9	85.4	80.6	82.4	82.1	77.6	83.6
OTPose [Jin *et al.*, 2022]	87.3	89.7	85.3	80.2	82.3	83.0	79.8	84.2
MTPose (Ours)	**89.4**	**92.4**	**90.1**	**87.3**	**85.7**	**89.7**	**88.1**	**89.0**

Table 1: Quantitative results on the PoseTrack2018 validation set.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Mean**
Tracktor++ w. poses [Bergmann *et al.*, 2019]	-	-	-	-	-	-	-	71.4
CorrTrack [Rafi *et al.*, 2020]	-	-	-	-	-	-	-	72.3
CorrTrack w. ReID [Rafi *et al.*, 2020]	-	-	-	-	-	-	-	72.7
Tracktor++ w. corr. [Bergmann *et al.*, 2019]	-	-	-	-	-	-	-	73.6
DCPose [Liu *et al.*, 2021]	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose [Liu *et al.*, 2022]	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
DiffPose [Feng *et al.*, 2023b]	84.7	85.6	83.6	80.8	81.4	83.5	80.0	82.9
TDMI-ST [Feng *et al.*, 2023a]	86.8	87.4	85.1	81.4	83.8	82.7	78.0	83.8
MTPose (Ours)	**92.0**	**91.7**	**88.7**	**85.5**	**86.4**	**86.6**	**85.3**	**88.3**

Table 2: Quantitative results on the PoseTrack21 validation set.

Results on PoseTrack21 We next evaluated MTPose on the PoseTrack21 validation set (Table 2). Our method achieved a significant improvement, reaching 88.3 mAP, which represents a notable 4.5 mAP gain over the existing state-of-the-art method, TDMI-ST [Feng *et al.*, 2023a]. However, unexpectedly, we observed the largest gains in the head joint on PoseTrack21, a joint typically less prone to movement. This result differs from the observations on PoseTrack2018. **Q. Why do the trends in the evaluation results differ between PoseTrack2018 and PoseTrack21?** A. The differences in evaluation trends between PoseTrack2018 and PoseTrack21 may be attributed to the distinct characteristics of the PoseTrack21 dataset. Specifically, the higher prevalence of small person poses in PoseTrack21 is likely a key factor. The nature of these poses often makes them less visible and more transient, presenting a challenge in effectively generating our motion-aware heatmaps. This aspect of the dataset may have contributed to the observed variations in performance trends.

Results on PoseTrack21 Motion Subset In response to this observed divergence, we conducted additional experiments using a specifically created PoseTrack21 Motion subset. This subset was compiled by selecting the top-5 videos with the largest average joint movements aiming to create a motion-intensive environment for evaluation. We computed the joint movements in the videos in the same way as we did in our previous experiments on PoseTrack2018. For comparison, we reproduced HRNet, DCPose, and FAMI-Pose [Sun *et al.*, 2019; Liu *et al.*, 2021; Liu *et al.*, 2022] using publicly available codes. Unfortunately, we were unable to reproduce DiffPose and TDMI-ST [Feng *et al.*, 2023a; Feng *et al.*, 2023b] as their codes were not publicly accessible. In comparison to those baselines, our MTPose's evaluation results on this motion subset further confirm the effectiveness of the proposed approach. We achieved an mAP of 83.4 in this Motion subset, a decrease of 4.9 mAP from the overall PoseTrack21 dataset results. In contrast, FAMI-Pose experienced a more pronounced drop of 7.8 mAP, decreasing from 81.2 mAP to 73.4 mAP. This outcome indicates that despite the challenges posed by PoseTrack21's dataset characteristics, MTPose remains robust in handling extensive motion scenarios. The Motion subset findings underscore our model's dynamic strengths, indicating that pronounced gains in less mobile joints, like the head in the full PoseTrack21 dataset, are influenced by the dataset's challenges rather than our approach's limitations.

Results on Sub-JHMDB Table 3 shows the performance of MTPose on the Sub-JHMDB dataset. Our model significantly outperforms other top-down heatmap regression methods, achieving an average improvement of 3.4 in the PCK. Notably, the advantage of MTPose is more evident at stricter PCK thresholds, demonstrating greater improvements as the threshold tightens from 0.2 to 0.1 and then 0.05. In direct comparison with DCPose, MTPose records substantial gains in average PCK, with increases of 4.0, 7.5, and 14.5, respectively. Moreover, MTPose demonstrates comparable performance to post-processing methods such as HANet, which rely on direct joint coordinate input for refinement [Zeng *et al.*, 2022; Jin *et al.*, 2023].

Strategy	Method	PCK@0.2								PCK@0.1	PCK@0.05
		Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Avg.**	**Avg.**	**Avg.**
Heatmap Regression	Thin-slicing Net [Song *et al.*, 2017]	97.1	95.7	87.5	81.6	98.0	92.7	89.8	92.1	-	-
	LSTM PM [Luo *et al.*, 2018]	98.2	96.5	89.6	86.0	98.7	95.6	90.0	93.6	-	-
	DKD [Nie *et al.*, 2019]	98.3	96.6	90.4	87.1	99.1	96.0	92.9	94.0	-	-
	SimpleBaseline [Xiao *et al.*, 2018]	97.5	97.8	91.1	86.0	99.6	96.8	92.6	94.4	81.3	56.9
	HRNet* [Sun *et al.*, 2019]	99.4	97.0	98.8	95.3	94.5	94.4	88.4	95.0	89.3	63.1
	DCPose* [Liu *et al.*, 2021]	99.4	97.3	99.1	95.3	95.3	94.4	90.0	95.4	90.0	67.8
	FAMI-Pose [Liu *et al.*, 2022]	99.3	98.6	94.5	91.7	99.2	91.8	95.4	96.0	-	-
	MTPose (Ours)	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**97.0**	**99.4**	**97.5**	**82.3**
+Post-Process	SimpleBaseline + DeciWatch [Zeng *et al.*, 2022]	99.8	99.5	**99.7**	**99.7**	98.7	99.4	96.5	98.8	94.1	79.4
	SimpleBaseline + HANet [Jin *et al.*, 2023]	**99.9**	**99.7**	**99.7**	**99.7**	**99.2**	**99.9**	**98.8**	**99.6**	**98.3**	**91.9**

Table 3: Quantitative Results on the Sub-JHMDB Dataset. Entries marked with "*" indicate reproduced results, as these methods were not originally evaluated on this dataset in their respective papers.

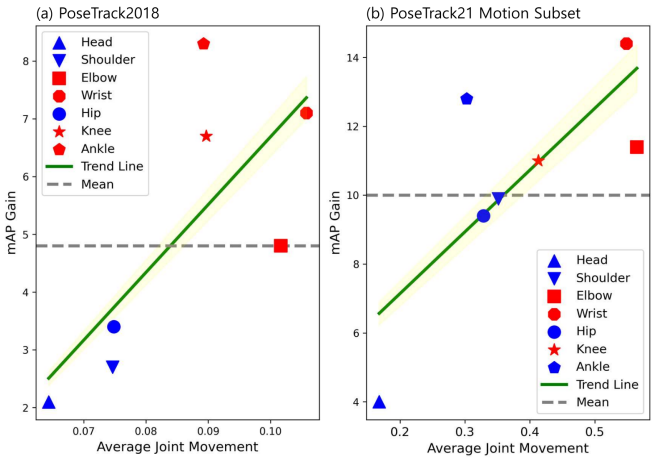

Figure 4: This figure depicts the mAP gains relative to average joint movement magnitude on the PoseTrack validation sets. Joints with significant motion, such as the elbow, wrist, and knee are shown in red, while less mobile joints like the head, shoulder, and hip are in blue. The green trendline highlights MTPose's significant performance improvements in joints with larger average movements.

Method	Head	Shou.	Elb.	Wri.	Hip	Knee	Ank.	**Mean**
HRNet [Sun *et al.*, 2019]	81.3	75.6	69.8	63.0	74.2	69.2	63.3	71.6
DCPose [Liu *et al.*, 2021]	83.4	76.9	69.5	65.1	75.5	69.8	65.6	73.0
FAMIPose [Liu *et al.*, 2021]	84.0	77.1	71.5	64.1	75.7	69.7	66.4	73.4
MTPose (Ours)	**88.0**	**87.0**	**82.9**	**78.5**	**85.1**	**80.7**	**79.2**	**83.4**

Table 4: Experiment on PoseTrack21 Motion Subset.

4.4 Ablation Study

To assess the impact of motion-aware heatmaps, we conducted an ablation study. This experiment compared the performance of original heatmaps (circular heatmaps with standard σ for both x and y axes) and Sigma-adjusted heatmaps (circular heatmaps with σ adjusted based on joint movement magnitude), to our Motion-aware heatmaps (elliptical heatmaps aligned with joint movement direction). Table 5 presents the results of this study. It reveals that adjusting the heatmap size according to motion magnitude does contribute to performance improvement. However, generating elliptical heatmaps that align with the direction of joint movement is particularly beneficial. This approach not only boosts overall performance but also significantly improves accuracy for joints that frequently move, such as the wrist and ankle.

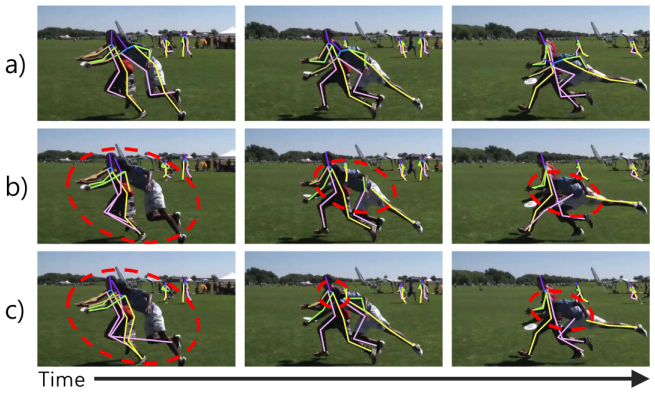

Figure 5: Qualitative results of our MTPose (a), FAMI-Pose (b), and DCPose (c) on PoseTrack21 Motion Subset.

Heatmaps	Adjust σ	Rotate θ	Wrist	Ankle	Mean
Original			86.1	86.8	87.8
Sigma-adjusted	✓		86.5 (+0.4)	87.2 (+0.4)	88.4 (+0.6)
Motion-aware	✓	✓	87.3 (+1.2)	88.1 (+1.3)	89.0 (+1.2)

Table 5: Ablation study on PoseTrack2018 validation set.

5 Conclusion

In this paper we propose a novel heatmap regression method to estimate 2D human poses in videos using the motion cues that capture the movements of joint keypoints in human motions. To do this, we introduce a new style of heatmaps what we call ***Motion-Aware Heatmaps*** based on the estimated joint movement to reflect the motion uncertainty of each joint point. In addition, we present a new architecture with the simple but effective architecture to integrate useful information in the joint motion cues to boost up the pose estimation performance. Through comprehensive experimental validation on three public benchmark datasets, our approach has demonstrated a notable improvement in pose estimation accuracy, particularly excelling in scenarios involving dynamic joint movements. The results underscore our method's effectiveness in predicting joints with significant motion and its robustness in various motion contexts.

Acknowledgments

This research was supported by the National Research Council of Science & Technology(NST) grant by the Korea government (MSIT) (No. CAP23081-000) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00254129, Graduate School of Metaverse Convergence (Sungkyunkwan University)).

References

[Bergmann *et al.*, 2019] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019.

[Bertasius *et al.*, 2019] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[Chen *et al.*, 2020] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding (CVIU)*, 2020.

[Cheng *et al.*, 2020] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[Dang *et al.*, 2022] Yonghao Dang, Jianqin Yin, and Shaojie Zhang. Relation-based associative joint location for human pose estimation in videos. *IEEE Transactions on Image Processing*, 2022.

[Doering *et al.*, 2022] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20963–20972, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fang *et al.*, 2017] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[Feng *et al.*, 2023a] Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, and Hyung Jin Chang. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17131–17141, 2023.

[Feng *et al.*, 2023b] Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, and Hyung Jin Chang. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14861–14872, 2023.

[Guo *et al.*, 2018] Hengkai Guo, Tang Tang, Guozhong Luo, Riwei Chen, Yongchen Lu, and Linfu Wen. Multi-domain pose network for multi-person pose estimation and tracking. In *The European Conference on Computer Vision (ECCV) Workshops*, 2018.

[Human-Machine-Interfaces-Trend-Report,] Human-Machine-Interfaces-Trend-Report. https://www.reply.com/en/topics/artificial-intelligence-and-machine-learning/human-machine-interfaces-trend-report.

[Iqbal *et al.*, 2017] Umar Iqbal, Anton Milan, and Juergen Gall. Posetrack: Joint multi-person pose estimation and tracking. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Jhuang *et al.*, 2013] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199, 2013.

[Jiang *et al.*, 2021] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021.

[Jin *et al.*, 2022] Kyung-Min Jin, Gun-Hee Lee, and Seong-Whan Lee. Otpose: Occlusion-aware transformer for pose estimation in sparsely-labeled videos. *arXiv preprint arXiv:2207.09725*, 2022.

[Jin *et al.*, 2023] Kyung-Min Jin, Byoung-Sung Lim, Gun-Hee Lee, Tae-Kyung Kang, and Seong-Whan Lee. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5725–5734, 2023.

[Li *et al.*, 2021] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[Liu *et al.*, 2021] Zhenguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. Deep dual consecutive network for human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[Liu *et al.*, 2022] Zhenguang Liu, Runyang Feng, Haoming Chen, Shuang Wu, Yixing Gao, Yunjun Gao, and Xiang Wang. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[Luo *et al.*, 2018] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[Luo *et al.*, 2021] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[Munea *et al.*, 2020] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The progress of human pose estimation: a survey and taxonomy of models applied in 2d human pose estimation. *IEEE Access*, 2020.

[Newell *et al.*, 2016] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *The European Conference on Computer Vision (ECCV)*. Springer, 2016.

[Nie *et al.*, 2019] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6942–6950, 2019.

[Rafi *et al.*, 2020] Umer Rafi, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 36–52. Springer, 2020.

[Song *et al.*, 2017] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4220–4229, 2017.

[Sun *et al.*, 2019] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[Tompson *et al.*, 2014] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[Tompson *et al.*, 2015] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Toshev and Szegedy, 2014] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[Wang *et al.*, 2020] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11088–11096, 2020.

[Xiao *et al.*, 2018] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *The European conference on computer vision (ECCV)*, 2018.

[Xu *et al.*, 2021] Lumin Xu, Yingda Guan, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Vipnas: Efficient video pose estimation via neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16072–16081, 2021.

[Xu *et al.*, 2022a] Xixia Xu, Yingguo Gao, Ke Yan, Xue Lin, and Qi Zou. Location-free human pose estimation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[Xu *et al.*, 2022b] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[Yang *et al.*, 2021] Yiding Yang, Zhou Ren, Haoxiang Li, Chunluan Zhou, Xinchao Wang, and Gang Hua. Learning dynamics via graph neural networks for human pose estimation and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8074–8084, 2021.

[Yu *et al.*, 2018] Dongdong Yu, Kai Su, Jia Sun, and Changhu Wang. Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network. In *Proceedings of the european conference on computer vision (ECCV) Workshops*, pages 0–0, 2018.

[Zeng *et al.*, 2022] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for $10\times$ efficient 2d and 3d pose estimation. In *European Conference on Computer Vision*, pages 607–624. Springer, 2022.

[Zhu *et al.*, 2019] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.