Fast and Accurate Pose Estimation in Videos based on Knowledge Distillation and Pose Propagation

Xiaomao Zhou
Center of Future Network Research
Purple Mountain Laboratories
Nanjing, China
xiaomaozhou26@gmail.com

Xiaolong Yu
Center of Future Network Research
Purple Mountain Laboratories
Nanjing, China
yuxiaolong@pmlabs.com.cn

Cheng Xu
Center of Future Network Research
Purple Mountain Laboratories
Nanjing, China
xucheng@pmlabs.com.cn

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892706

***Abstract*—Existing video-based human pose estimation methods typically adopt large networks to perform body joints localization on all frames. Despite of impressive accuracy performance, the relatively high memory and computation requirements significantly burden their applicability on resource-constraint systems (e.g., embedded devices). To solve this issue, this paper proposes a novel yet effective lightweight framework, called FVPE, for fast and accurate human pose estimation in videos. Specifically, FVPE adopts the knowledge distillation (KD) strategy to train a small pose estimator network, which is capable of executing rapidly with low computational cost. To increase the overall efficiency, FVPE exploits the temporal coherence between successive video frames and explicitly propagates body joints from previous frames rather than naively extracting them using a pose estimator. Furthermore, FVPE introduces an online key-frame selection scheme to decide whether the current pose should be calculated by the pose estimator or be propagated from the previous key-frame, being able to flexibly deal with video sequences with different length, frame rate, pose complexity, etc.. Experiments on Penn Action and Sub-JHMDB datasets demonstrate that the proposed method achieves comparative accuracy, but with substantial speed-up.**

I. Introduction

Video human pose estimation aims to recognize and localize human anatomical joints in each frame and serves as a significant basis for several vision tasks such as action recognition, virtual reality, and human-computer interaction. Recent advances in Convolutional Neural Networks (CNNs) [1] have significantly boosted the accuracy of pose estimation models. However, the strong performance is typically achieved by training and deploying resource-intensive networks with large depth and width [2], which significantly weakens their inference efficiency and restrains their applicability and scalability in realistic applications. For example, large neural networks are usually challenging to be employed on resource-limited devices, e. g., smartphones, mobile robots.

Considering that accuracy and efficiency are equivalently critical, recent efforts have been devoted to designing pose estimation models with small model size, light computation cost, and high estimation accuracy. However, most lightweight networks can hardly produce satisfying pose estimation results due to their low representational capacity nature. To solve this problem, a promising solution is knowledge distillation [3] which adopts the teacher-student paradigm and enables the small networks (student) to mimic the large well-trained networks (teacher). The idea behind KD is to use the soft probabilities from a trained teacher as supplementary information to promote the student learning outcomes. It has been verified valid and widely adopted in many vision tasks including, image classification [4], object detection [5], and semantic segmentation [6].

In addition to resorting to light-weight pose estimation models, the overall computational costs can be significantly reduced by mitigating the inference redundancy, i. e., selectively passing several keyframes through the pose estimation model and propagating poses of other frames from their neighboring keyframe, instead of feeding all frames into the pose estimation model. Such an idea is based on the fact that consecutive frames in videos share great contextual and geometric consistency and it is possible to obtain the current pose by transforming previous ones according to relative joint motion offsets. As such, most of the body joint localization tasks are converted to perform pose propagation across frames, which can be easily handled by a lightweight network. Based on this strategy, the video pose estimation efficiency can be significantly improved without performance degradation. Nevertheless, most relevant works select keyframes using cumbersome RNNs [7] or manually designed policies [8], either bringing in extra computation burdens or failing to deal with video sequences with various framerate and complexity.

In this paper, we focus on improving the efficiency of human pose estimation in videos while preserving comparable accuracy. To this end, we propose the FVPE framework, which intensively leverages KD and the pose propagation mechanism. Specifically, FVPE first adopts KD to train a lightweight network to perform image-based pose estimation with high efficiency. Then, based on the pose propagation mechanism, FVPE intelligently converts most computationally expensive pose regression calculations into computationally friendly pose propagation tasks, further improving the inference efficiency. Furthermore, an online key-frame selection (OKFS) module is proposed to select informative frames in real time, based on which the incoming pose should be calculated by the pose estimator or be propagated from the previous key-frame. OKFS is a lightweight network and adopts KD to learn from a complex well-trained teacher model which can

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

accurately predict neighboring frames. Compared to existing works, FVPE can accurately and efficiently localize body joints in all frames, while being robust to video sequences with different framerate and complexity. Extensive experiments on two widely used benchmarks Penn Action and Sub-JHMDB demonstrate the efficiency and effectiveness of the proposed FVPE for resolving human pose estimation in videos.

The main contributions of this work can be summarized as follows:

- We propose a novel framework, FVPE, to facilitate light-weight networks in video pose estimation with high accuracy and efficiency, by leveraging KD to compress the model size and the pose propagation mechanism to mitigate the inference redundancy, promoting deep pose estimation methods in real applications.
- We introduce a simple yet efficient online key-frame selection (OKFS) network to detect informative frames in videos, based on which to perform pose propagation, avoiding unnecessary processing of non-key frames.
- We conduct extensive experiments on two benchmarks to demonstrate the effectiveness of the proposed framework, showing comparable accuracy and outperforming efficiency.

II. Related Work

A. Human Pose Estimation

Existing CNN-based methods have facilitated human pose estimation with superior performance, which usually directly regress joint coordinates [10] or joint confidence heatmaps [11] using powerful neural networks and adopts either strong backbones [12] or multi-stage refinement architectures [13] in order to achieve competitive performance. Considering inference efficiency, recent works have also studied approaches to promoting applicability and scalability in resource-constraint systems. For example, some works study on designing lightweight backbone networks based on concepts including depthwise convolution [14], network binarization [15], attention mechanism [16], etc. Zhang et al. [17] adopt neural architecture search (NAS) to find optimal backbone networks in pose estimation. To mitigate the performance drop caused by low-capacity networks, Yang et al. [18] propose to jointly generate body joints by considering their structural relationships. Zhang et al. [19] present a KD-based training method to transfer the pose structure knowledge from a teacher network, whose idea has been followed by many works.

For pose estimation in videos, most works focus on modeling temporal cues to boost the performance, where the optical flow [20], 3D CNN [21], and recurrent neural networks (RNNs) [22] are widely adopted. Furthermore, Huang et al. [23] consider the temporal consistency between consecutive frames and propose to directly propagate poses from the well-estimated neighboring ones. Nie et al. [24] use a lightweight distillator to online distill pose kernels via leveraging temporal cues from the previous frame. Yerule et al. [25] research on the correlation of each individual pose between consecutive frames and present a dynamicity-based mechanism to adaptively perform cropping, dropping or pose estimation. In this work, we follow the same strategy and introduce the OKFS to intelligently decide when to perform pose estimation or pose propagation.

B. Knowledge Distillation

KD has been a widely-adopted solution to train lightweight networks with a strong performance by transferring knowledge from large teacher networks, where the distilled targets include logits outputs [19], intermediate features [26], or structural relationships [26]. Early works adopt a two-stage training schema to perform one-way KD, which gradually derives many variants. Li et al. [27] introduce an online KD framework for efficient pose estimation based on a multi-branch architecture, without the need for pre-trained teacher models. Guo et al. [28] enable knowledge to be transferred among different agents via collaborative learning. Weinzaepfel et al. [29] propose to distill knowledge from multiple expert models in different fields into one single student network. Dou et al. [30] exploit to perform KD between different modalities. In this work, we train the compact pose estimator and the OKFS network based on the KD strategy.

III. Proposed Method

The framework of the proposed SFVP is illustrated in Fig. 1, which consists of three components: the KD-based light-weight pose estimation (KD-PE), the online key frame selection module (OKFS), and the pose propagation network (PPNet). we start with an introduction and then describe each component with implementation details.

A. Overall Pipeline

Given a video sequence with T consecutive frames $\{I_t\}_1^T$, where $I_t \in \mathbb{R}^{H\times W\times 3}$denotes the frame at time step t, the objective of FVPE is to generate frame-wise pose heatmaps $\{h_t\}_1^T$ with high accuracy and efficiency, where $h_t \in \mathbb{R}^{H/S\times W/S\times M}$ and W, H, S, M denote the scale parameters. For this aim, the KD-PE is trained to perform fast pose estimation and works only on keyframes $\{I^m\}_1^M$ which are selected by the OKFS, where the first frame is naturally regarded as the keyframe. For frame $I_t, (t \in [2, T])$, it is first fed into the OKFS to decide whether it should be selected as a keyframe or not, based on the pixel difference between I_t and its preceding keyframe I^m. If I_t is selected as a keyframe, its pose will be estimated by the KD-PE. If not, its pose will be generated by the PPNet which performs pose propagation from I^m, whose computation costs are much smaller than the KD-PE. In this way, the overall computation costs can be greatly reduced since the pose calculation of the majority of frames can be simplified, while the accuracy can also be maintained since the KD-PE will be in-time adopted when the current poses can not be propagated accurately.

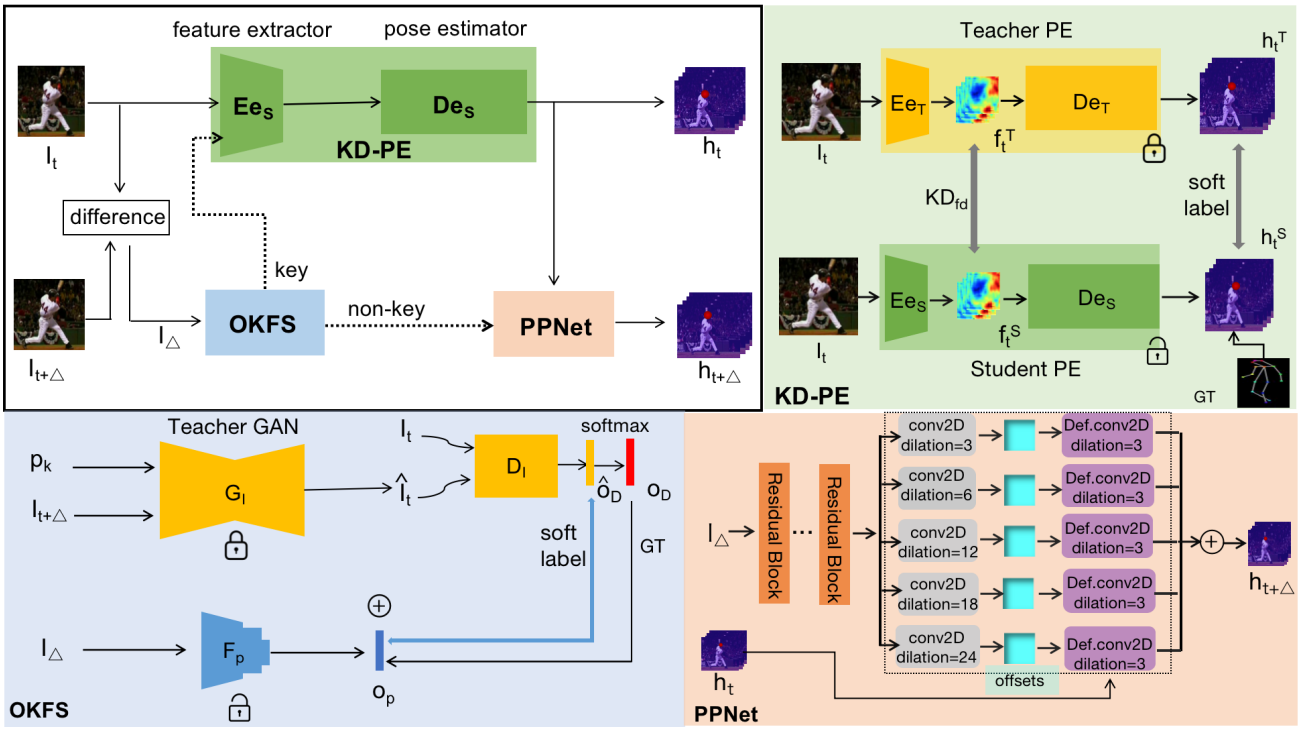

Fig. 1. Overview of the proposed FVPE. It consists of a KD-based pose estimation network(KD-PE), an online keyframe selection module(OKFS), and a pose propagation network(PPNet). The KD-PE is used to calculate poses of the keyframes and PPNet is used to calculate poses of the non-key frames, where the OKFS is responsible for deciding whether an incoming frame should be selected as key or not.

B. Compact KD-based Pose Estimator

KD-PE is a light CNN network and adopts KD to transfer relevant knowledge from a powerful yet cumbersome pose estimation network for better training. Specifically, we follow the designs in [19] and build the Teacher PE with 8 stacked hourglass modules, where each one consists of 9 Residual blocks with 256 channels within each layer. While the KD-PE (the Student PE) consists of 4 stacked hourglass modules and each module has 128 channels, whose computational cost is only one-sixth of the Teacher PE. Different from the conventional KD strategy which only uses the outputs of the teacher network as complementary supervision to train the student network, we propose an integral KD approach that includes logit distillation KD_{soft} and feature distillation KD_{fd}. Concretely, KD_{soft} regards the output of the teacher network as the soft label, integrating with the ground truth label, to train the KD-PE. KD_{fd} ensures the high-level features of the KD-PE to be consistent with the teacher network. The loss function LD_{soft} is defined as follows:

$$LD_{PE}^{soft} = \frac{1}{K}\sum_{k=1}^{K}||h_k^T - h_k^S||_2^2 \tag{1}$$

where h_k^T and h_k^S represent the confidence maps of the $k-th$ joint predicted by the teacher network and the student network, respectively. Here, the L_2 loss is chosen to maximise the comparability between LD_{soft} and the original pose loss.

The loss function LD_{fd} is defined as follows:

$$LD_{PE}^{fd} = \frac{1}{(W' \times H')^2}\sum_{i\in R}\sum_{j\in R}(f_{ij}^T - f_{ij}^S)^2 \tag{2}$$

where f_{ij} denotes the similarity between the i-th pixel and j-th pixel. W' and H' represent the width and height of the feature map, and $R = 1, 2, ...W' \times H'$ denotes all the pixels. f_{ij} is computed from the features f_i and f_j as:

$$f_{ij} = f_i^T f_j / (||f_i||_2 ||f_j||_2) \tag{3}$$

Therefore, the overall KD-PE loss function, which combines the MSE (Mean-Squared Error) loss L_{MSE}^{PE} using the ground truth label with LD_{soft}^{PE} and LD_{fd}^{PE}, can be formulated as:

$$L_{PE} = L_{PE}^{MSE} + \lambda_1 LD_{PE}^{soft} + \lambda_2 LD_{PE}^{fd} \tag{4}$$

where λ_1 and λ_2 are the hyper-parameters that modulate the proportion of different losses, which are set as 0.3 and 0.2 according to our experimental results.

C. Online Key Frame Selection Module

The OKFS is to decide if an incoming frame should be selected as a keyframe or not. It exploits the similarity among consecutive frames and is trained to make the correct decision based on the pixel difference $f_\triangle$ between the incoming frame $I_{m+\triangle}$ and its preceding keyframe I_m. To fulfill this, we also adopt KD for better training, where the teacher network G_I is a GAN-based image generation network and the student

network F_p is a lightweight CNN network performing simple classification. G_I is able to generate the target image I_m given the current image $I_{m+\triangle}$ and the desired pose p_m, but only maintains good performance among images with similar appearance and geometry features, i. e. within neighboring frames in videos. D_I is able to distinguish between the image $\hat{I}_m$ generated by the G_I and its ground truth I_m. Our aim is to transfer the knowledge of G_I and D_I to F_p so that it is able to make the same decision as D_I though the input is given in a different fashion.

For implementation details, G_I and D_I are built with large networks, whose structures follow the designs in PG^2 [31], and are well pre-trained on relevant datasets. F_p is a light network consisting of 3 depth-wise convolutional layers followed by one fully connected layer. During training, the inputs to G_I are the current image $I_{m+\triangle}$ and the pose p_m of the preceding keyframe I_m, its output is the generated image $\hat{I}_m$ which is then fed into the discriminator D_I. The outputs of D_I include $\hat{o}_D$ (before softmax) and o_D, which are regarded as the soft label and the ground truth respectively, to stimulate F_p training. The loss function of F_p is described as:

$$L_{KS} = \alpha||o_p - o_D||_2 + (1-\alpha)||o_p - \hat{o}_D||_2 \tag{5}$$

where the first part is the MSE loss to the ground truth and the second part is the distillation loss. α is the balancing factor between these two losses, which is empirically set as 0.5.

D. Light-weight Pose Propagation Network

While KD-PE is dedicated to estimating poses of keyframes, PPNet is used to generate poses of those non-key frames, which account for the majority of frames in videos. Similar to the paradigm in [23], we propose to use a light-weight network to directly perform pose propagation among neighboring frames, i. e., the current pose $p_{m+\triangle}$ can be propagated from the previous one p_m according to the joint offsets between the two frames. The network structure of PPNet is shown in Fig.1, which consists of a stack of 3×3 residual blocks followed by 5 3×3 convolutional layers with different dilation rates $d \in \{3, 6, 12, 18, 24\}$ to predict five sets of offsets at all pixel locations. Then, the predicted offset tensors are used to transform pose p_m via deformable convolutions, outputting the propagated pose $p_{m+\triangle}$. During training, the MSE based loss is used to optimize PPNet via back-propagation, which is formulated as:

$$L_{PP} = \frac{1}{K}\sum_{k=1}||h_m - h_m^{gt}||_2^2 \tag{6}$$

where p_k and p_k^{gt} denote the predicted pose and the ground truth heatmaps, respectively.

Different from existing works that usually perform pose propagation with a fixed keyframe interval, i. e., selecting one keyframe every k frames, which fails to fully exploit the dynamic correlation between neighboring frames. In this work, we use the OKFS to dynamically select the keyframes, which mainly brings in three advantages: (1) Minimizing the overall computational burdens while maintaining the performance. Since only a small set of keyframes are fed into the KD-PE, the majority of pose calculations are done by the PPNet, leading to less computation. (2) Being robust to video sequences with different configurations including range, framerate, complexity, etc., due to the adaptive selection frequency of keyframe. (3) Possessing high tolerance to the error of OKFS. When the OKFS mistakenly recognizes a non-key frame as a keyframe, this only affects the pose calculation of that single frame, while the calculations of its precedent and subsequent frames are almost unaffected. When a keyframe is missed, one of its subsequent frames is very likely to be selected as the keyframe. We can also adjust the evaluation criterion of OKFS to balance the accuracy and efficiency of the proposed framework.

IV. Experiments

In this section, we elaborate on the implementation details of the proposed FVPE and evaluate it with qualitative and quantitative experiment results on different datasets.

A. Experimental Setup

Datasets The proposed FVPE is evaluated on two widely used benchmarks: Penn Action [32] and Sub-JHMDB [33]. Penn Action is a large-scale unconstrained dataset and contains 2326 video sequences distributed over 15 actions, of which 1258 clips are used for training and 1068 clips are for testing. The annotations include the 2D person bounding boxes, coordinates, and visibility of 13 body joints, where only the visible joints are involved in the evaluation. In contrast, Sub-JHMDB consists of 316 video clips with 11200 frames, where 15 visible joints of each complete body are annotated. Following previous works, the training and testing samples are split with three different schemes, and we separately conduct evaluations on these three splits and report the average precision.

Training Details FVPE adopts a multiple-step training schema, where the KD-PE, the OKFS, and the PPNet are trained independently. The training is accelerated by using a NVIDIA T4 GPU.

For training the KD-PE, we use the original Hourglass as the teacher pose estimator and the KD-PE with the customized Hourglass architecture as the student pose estimator, where the teacher model is first pre-trained on the MPII dataset [34]. The KD-PE is fine-tuned on the Penn Action dataset and Sub-JHMDB dataset for 70 and 40 epochs, respectively. RMSProp optimizer is adopted with 10^{-5} weight decay and the initial learning rate decreases linearly from 2.5×10^{-4} to 0. During training and inference, all the images are resized to 256×256 based on the center and scale of the person instance, where data augmentations including random scaling, rotation, and flipping are also conducted.

We then adopt the KD strategy to train the OKFS, where the teacher Generator and Discriminator are firstly trained on the Penn Action dataset and Sub-JHMDB dataset. During training, we select a frame from a video clip and then randomly select a neighboring frame within 10 frame intervals (precedent or subsequent) as the target. The selected two frames are fed into the teacher GAN, whose outputs are used to train the

Fig. 2. Qualitative results of the proposed FVPE. Yellow bounding boxes indicate the keyframes chosen by the OKFS. The red skeletons are the ground truth, the green skeletons are predicted by the KD-PE, and the yellow skeletons are calculated by the PPNet.

network F_p. Adam optimizer with a learning rate of 0.0001 is adopted and the total epoch is 10k. The aforementioned image processing strategies are also performed.

We then train the PPNet following similar procedures in [23]. Two frames are first selected using the same policy in the training of OKFS and their pixel difference is then fed into the PPNet to output the target pose. The loss to the ground truth pose is back-propagated to update the parameters of PPNet. We use the Adam optimizer with a base learning rate of 0.0001, which decreases linearly to 0 after 200k iterations. We set the batch size as 8 and perform the seven-scale inference during evaluation.

Evaluation Metrics We use the PCK metric and the frame per second (FPS) to evaluate the pose estimation accuracy and inference efficiency, respectively. PCK reports the percentage of estimated keypoints lying within a normalized distance to the ground truth. While FPS represents the inference speed of the whole framework.

B. Experimental Evaluations

Qualitative Results Fig 4 presents qualitative results on different datasets. It can be seen that the proposed FVPE first proposes a certain number of keyframes in videos and then is able to generate poses of other non-key frames by performing propagation from poses of these keyframes. Visually, the generated poses demonstrate high consistency with the ground truths, showing that our FVPE can perform accurate pose estimation. Furthermore, our model preserves high accuracy in poses with unusual postures and severe self-occlusions since it intelligently exploits the temporal coherence between successive frames.

Quantitative Results We also present quantitative results using metrics described above on both datasets and compare our FVPE against several recent methods, including LSTM PM [7], Pose Wrapper [35], HRNet [12], DKD [24], PPN [19], to verify its effectiveness from two aspects, i.e, accuracy, and efficiency. As shown in Table 1, the proposed FVPE demonstrates comparable PCK@0.2 scores to state-of-the-art methods, with only a little performance degradation. Table 2 reports the results of different methods in performing frame-wise pose estimation on a certain number of video clips, which consist of 2183 frames. As presented, FVPE achieves the fastest inference speed (38.4FPS), while using the fewest parameters (11.3M) and the smallest average computation FLOPs (6.7G). Specifically, its inference speed is about $3\times$ faster than the best performer (HRNet) and requires only 16.2% average computation FLOPs. Even compared to the most efficient alternative competitor (PPN), its parameters and average computation FLOPs also decrease by 67.6% and 11.8%, respectively, gaining 26.3% improvement in inference speed. These results clearly demonstrate the superior accuracy and efficiency of our model for pose estimation in videos.

TABLE I
QUANTITATIVE COMPARISON WITH OTHER METHODS USING PCK@0.2 ON PENN ACTION AND SUB-JHMDB DATASETS.

Datasets	Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	PCK
Penn Action	LSTM PM	98.9	98.6	96.6	96.6	98.2	98.2	97.5	97.7
	Pose Wrapper	98.3	98.2	95.5	96.7	98.1	98.0	96.8	97.4
	HRNet	98.7	98.6	98.1	98.4	98.5	98.6	98.8	98.7
	DKD(ResNet-50)	98.8	98.7	96.8	97.0	98.2	98.1	97.2	97.8
	PPN(ResNet-50)	99.2	98.9	97.8	97.4	98.7	98.5	98.3	98.7
	Ours	98.6	98.3	97.4	97.9	98.0	98.4	98.1	98.1
Sub-JHMDB	LSTM PM	98.2	96.5	89.6	86.0	98.7	95.6	90.9	93.6
	Pose Wrapper	97.9	96.8	90.4	86.3	98.4	96.4	91.7	93.9
	HRNet	98.4	97.2	93.8	93.1	99.2	97.7	95.3	96.5
	DKD(ResNet-50)	98.3	96.6	90.4	87.1	99.1	96.0	92.9	94.0
	PPN(ResNet-50)	99.0	98.3	92.5	90.9	99.4	98.3	95.0	96.4
	Ours	98.4	98.2	91.9	91.2	98.7	97.5	93.9	95.6

TABLE II
INFERENCE COMPLEXITY COMPARISONS OF DIFFERENT MODELS ON THE TEST VIDEO CLIPS. FLOPS, PARAMS, AND FPS ARE AVERAGED OVER THE TOTAL FRAMES. THE INFERENCE IS CONDUCTED ON A LAPTOP WITH AN INTEL CORE I5-5300 CPU.

Methods	FLOPs(G)	Params(M)	Time(s)	FPS
LSTM PM	227.9	271.5	178.7	12.2
Pose Warper	28.4	11.7	105.6	20.7
HRNet	34.6	77.5	186.6	11.7
DKD	8.9	38.7	156.84	13.9
PPN	7.6	34.3	77.6	28.3
Ours	**6.7**	**11.3**	**56.9**	**38.4**

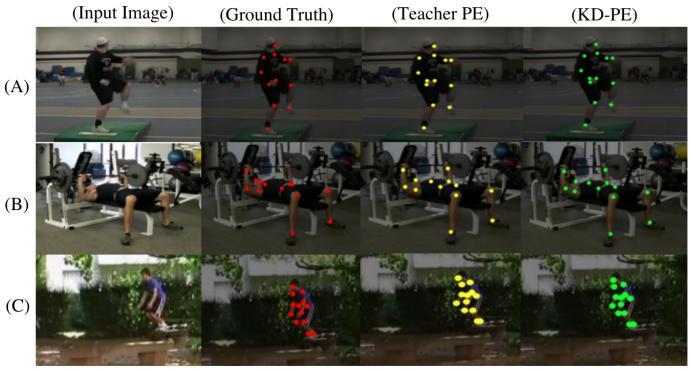

Fig. 3. Pose estimation results of the proposed KD-PE. From left to right are: the input images, the ground truth joint confidence heatmaps, the outputs of the Teacher PE, and the outputs of the KD-PE. From top to down are: (A) missing labels recovered by the Teacher PE; (B) error labels corrected by the Teacher PE; (C) ground truth labels enhanced by the Teacher PE, where the joint confidence maps are enlarged.

C. Performance Analysis

To delve into our FVPE framework, we explicitly evaluate the learning results of its main components, i. e., the KD-Pe and the OKFS, and conduct the following experiments.

KD Learning We test the effectiveness of using KD strategies on the lightweight PE and OKFS. For this aim, we simply disable and enable the KD component, i.e., using or not using supervisions from the teacher network, in the training procedure and then compare the performance on different datasets. As shown in Table 3, the KD brings in 10.53% PCK accuracy boost on the PE and 5.77% accuracy improvement on the OKFS, respectively. To further reveal the underlying reasons, we visualize the pose estimation results of several examples with and without KD in Fig. 3. We can observe that KD can effectively promote the student PE to deal with challenging cases including missing annotations (Fig. 3 (A)), error labeling(Fig. 3 (B)), and hard training images(Fig. 3 (C)), showing that the distilled knowledge from the trained teacher network reflects more about the data than its true label.

TABLE III
THE PERFORMANCE COMPARISONS OF THE KD-PE AND OKFS WITH AND WITHOUT KD LEARNING.

Methods	wo. KD	wt. KD
KD-PE (PCK)	90.1	95.3
OKFS (Accuracy)	89.3	98.7

Analysis of the OKFS We first visualize the qualitative results of the trained Teacher GAN network. As shown in Fig. 4, the generated non-key-frame images preserve consistent global structures with the corresponding ground truth images. Although the appearance details, e.g., clothing textures, facial expressions, and background objects, are sometimes distorted or missing, this brings in ignorant effects on the pose estimation results. As for the keyframes (highlighted by the yellow boxes), the generated images always possess huge differences, e. g., missing body parts, distorted actions, and blurred appearances, with the ground truth image. The OKFS can select the keyframes based on the pixel difference between the generated image and its ground truth.

We then evaluate the accuracy of the OKFS on the final performance. For this aim, we compare the performances of using different OKFS thresholds, i. e., adjusting the evaluation criticism of selecting the keyframes. As presented in Table 4, different thresholds result in a different number of keyframes to be selected, where the accuracy and inference efficiency change significantly. For example, when setting the threshold to be high, fewer keyframes will be selected, the average inference speed will increase, but the accuracy will decrease. When setting the threshold to be low, more frames will be selected as the keyframes, but the overall

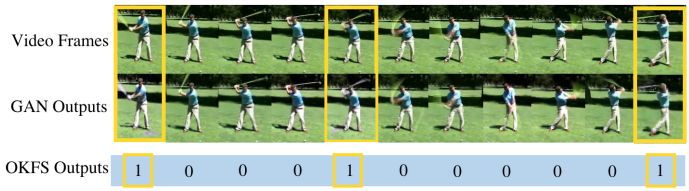

Fig. 4. Qualitative results of the Teacher GAN. The OKFS is able to select the keyframes based on the pixel difference of the generated images and the ground truth images.

TABLE IV
PERFORMANCE COMPARISONS OF USING DIFFERENT OKFS THRESHOLDS.

Thresholds	PCK	Keyframes	Time(s)	FPS
0	**95.8**	2183	152.7	14.3
Low (0.23)	95.5	1259	108.1	20.2
Medium (0.67)	95.4	261	77.6	38.4
High (0.86)	87.7	139	**47.5**	**45.6**

accuracy and inference speed make negligible changes. This is because the poses of keyframes are calculate by the KD-PE, whose computation costs are larger than the PPNet, thus more keyframes result in inference efficiency decrease. Meanwhile, although the PPNet brings in efficiency increase, it can only maintain its performance within a limited frame range, thus too few keyframes result in the overall accuracy drop. In the experiments, we empirically set the threshold to be 0.67, aiming for the balance between accuracy and efficiency.

D. Ablation Studies

We further conduct a series of ablation studies to analyze the contribution of each component on the overall performance, where several different models are designed to perform frame-wise pose estimation. These models consist of : 1) **Baseline** that adopts the student PE architecture and only uses the truth label for training; 2) **Baseline+KD** that adopts the student PE architecture and use the KD strategy for training; 3) **Baseline+KD+PPNet+Fixed Keyframe Selection (FKFS)** that calculates the poses of keyframes and non-key frames using KD-PE and PPNet, respectively, where keyframes are selected every 5 frames; 4) **Baseline+KD+PPNet+OKFS (Ours)** that adopts the proposed pipeline, where keyframes are selected by the OKFS.

Table 5 presents the quantitative results of the ablation studies. We can observe that the KD-PE demonstrates comparable accuracy with the Teacher PE, outperforming the **Baseline** by 1.4%, which demonstrates the effectiveness of the KD in promoting training. Meanwhile, the adoption of PPNet significantly improves the inference efficiency. The overall inference speed of the **Baseline+KD+PPNet+FKFS** model and **Ours** model increase by 20.0% and 34.9%, respectively. However, the accuracy performance of the former model is inferior to the latter one since it uses a keyframe selection policy with fixed intervals and can not deal with situations where the selection frequency of keyframes needs to be adjusted. In contrast, the OKFS is able to dynamically propose keyframes according to the pose complexities and the frame rate of videos, thus being able to fully exploit the dynamic correlations between consecutive frames in videos.

TABLE V
COMPARISON RESULTS OF THE ABLATION STUDIES.

Models	PCK	Keyframes	Time(s)	FPS
Teacher PE	**95.7**	2183	177.5	12.3
Baseline	94.1	2183	97.0	22.5
+KD	95.5	2183	97.0	22.5
+OKFS	95.5	276	97.0	22.5
+KD+PPNet+FKFS	94.3	374	77.6	34.6
+KD+PPNet+OKFS	**95.6**	261	**63.1**	**38.4**

V. CONCLUSION

In this paper, we propose FVPE, a novel yet efficient framework for pose estimation in videos, which adopts the KD strategy and the pose propagation mechanism to promote the overall accuracy and efficiency. Specifically, FVPE first improves the single-image based pose estimation by training a compact pose estimation network, which adopts an integral KD strategy. To further reduce the overall computation burdens, FVPE exploits the temporal coherence among consecutive videos frames and performs pose propagation using a lightweight network, rather than resorting to the pose estimator, which is able to obtain the poses in a more computationally-friend way. Additionally, FVPE also introduces a KD-based keyframe selection network to online select keyframes in videos, which are fed into the pose estimator and then used to propagate the poses of the non-key frames. In this way, FVPE enables lightweight networks to perform accurate and fast pose estimation, while maintaining satisfactory performance in videos with different length, framerate and pose complexity, etc. Extensive experimental results on the PEN Action and Sub-JHMDB datasets verify the effectiveness and efficiency of our proposed model.

REFERENCES

[1] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," Progress in Artificial Intelligence, 9(2), pp.85-112, 2020.
[2] B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5386-5395, 2020.
[3] G. Hinton, O. Vinyals, J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[4] M. Huang, Y. You, Z. Chen, Y. Qian, K. Yu, "Knowledge Distillation for Sequence Model," In Interspeech, pp. 3703-3707, 2018.
[5] G. Chen, W. Choi, X. Yu, T. Han, M. Chandraker, "Learning efficient object detection models with knowledge distillation," Advances in Neural Information Processing Systems, 30, 2017.
[6] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, "Structured knowledge distillation for semantic segmentation," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2604-2613, 2019.
[7] Y. Luo, J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, L. Lin, "Lstm pose machines," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5207-5215, 2018.
[8] Y. Zhang, Y. Wang, O. Camps, M. Sznaier, "Key Frame Proposal Network for Efficient Pose Estimation in Videos," In European Conference on Computer Vision, pp. 609-625, 2020.

[9] A. Newell, K. Yang, J. Deng, "Stacked hourglass networks for human pose estimation," In European conference on computer vision, pp. 483-499, 2016.
[10] X. Sun, B. Xiao, F. Wei, S. Liang, Y. Wei, "Integral human pose regression," In Proceedings of the European Conference on Computer Vision, pp. 529-545, 2018.
[11] A. Bulat, G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," In European Conference on Computer Vision, pp. 717-732, 2016.
[12] K. Sun, B. Xiao, D. Liu, J. Wang, "Deep high-resolution representation learning for human pose estimation," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693-5703, 2019.
[13] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, J. Sun, "Rethinking on multi-stage networks for human pose estimation," arXiv preprint arXiv:1901.00148, 2019.
[14] Z. Zhang, J. Tang, G. Wu, "Simple and lightweight human pose estimation," arXiv preprint arXiv:1911.10346, 2019.
[15] A. Bulat, G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," In Proceedings of the IEEE International Conference on Computer Vision, pp. 3706-3714, 2017.
[16] G. Ning, Z. Zhang, Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," IEEE Transactions on Multimedia, 20(5), pp. 1246-1259, 2017.
[17] W. Zhang, J. Fang, X. Wang, W. Liu, "Efficientpose: Efficient human pose estimation with neural architecture search," Computational Visual Media, 7(3), pp.335-347, 2021.
[18] Y. Yang, J. Yin, "Relation-Based Associative Joint Location for Human Pose Estimation in Videos," arXiv preprint arXiv:2107.03591, 2021.
[19] F. Zhang, X. Zhu, M. Ye, "Fast human pose estimation," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3517-3526, 2019.
[20] T. Pfister, J. Charles, A. Zisserman, "Flowing convnets for human pose estimation in videos," In Proceedings of the IEEE international conference on computer vision, pp. 1913-1921, 2015.
[21] L. Ge, H. Liang, J. Yuan, D. Thalmann, "3d convolutional neural networks for efficient and robust hand pose estimation from single depth images," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1991-2000, 2017.
[22] Z. Hu, Y. Hu, J. Liu, B. Wu, D. Han, T. Kurfess, "A CRNN module for hand pose estimation," Neurocomputing, 333, pp.157-168, 2019.
[23] X. Huang, W. Deng, H. Shen, X. Zhang, J. Ye, "PropagationNet: Propagate Points to Curve to Learn Structure Information," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7265-7274, 2020.
[24] X. Nie, Y. Li, L. Luo, N. Zhang, J. Feng, "Dynamic kernel distillation for efficient pose estimation in videos," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6942-6950, 2019.
[25] S. N. Yerule, Y. F. Wu, C. C. Kao, Y. C. Tseng, "Dynamicity-based Crop-Drop: A Context-based 2D Pose Refreshing Algorithm," In 2020 International Conference on Pervasive Artificial Intelligence, pp. 258-263, 2020.
[26] X. Xu, Q. Zou, X. Lin, Y. Huang, Y. Tian, "Integral knowledge distillation for multi-person pose estimation," IEEE Signal Processing Letters, 27, pp.436-440, 2020.
[27] Z. Li, J. Ye, M. Song, Y. Huang, Z. Pan, "Online Knowledge Distillation for Efficient Pose Estimation," In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11740-11750, 2021.
[28] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, P. Luo, "Online knowledge distillation via collaborative learning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11020-11029, 2020.
[29] P. Weinzaepfel, R. Brégier, H. Combaluzier, V. Leroy, G. Rogez, "Dope: Distillation of part experts for whole-body 3d pose estimation in the wild," In European Conference on Computer Vision, pp. 380-397, 2020.
[30] Q. Dou, Q. Liu, P. A. Heng, B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation," IEEE Transactions on Medical Imaging, 39(7), pp.2415-2425, 2020.
[31] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, "Pose guided person image generation," arXiv preprint arXiv:1705.09368, 2017.
[32] W. Zhang, M. Zhu, K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," In Proceedings of the IEEE International Conference on Computer Vision, pp. 2248-2255, 2013.
[33] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M. J. Black, "Towards understanding action recognition," In Proceedings of the IEEE International Conference on Computer Vision, pp. 3192-3199, 2013.
[34] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp. 3686-3693, 2014.
[35] G. Bertasius, C. Feichtenhofer, D. Tran, J. Shi, L. Torresani, "Learning temporal pose estimation from sparsely-labeled videos," arXiv preprint arXiv:1906.04016, 2019.