Temporal-Interim Pose Synthesis and Distillation for Dynamic Human Pose Estimation

Renjie Zhang, Di Lin, *Member, IEEE*, Xin Wang, Ruonan Liu, *Senior Member, IEEE*, Bin Sheng, *Member, IEEE*, George Baciu, *Senior Member, IEEE*, C. L. Philip Chen, *Life Fellow, IEEE*, and Ping Li, *Member, IEEE*

***Abstract*—In the task of dynamic human pose estimation, the temporal relationships between human body parts should be captured comprehensively to understand the dynamic human motions, where the correlated motion information eventually helps to recognize body parts. The popular methods are successful in terms of utilizing long-term motion information captured by low-speed cameras. Yet, they neglect the underlying intermediate motions between captured frames, which comprise the temporal-interim poses lost in the video. In this paper, we introduce a novel framework, Temporal-interim Pose Synthesis and Distillation, to produce and leverage the intermediate motion information for dynamic motion establishment. The pose synthesis yields the visual feature maps of the intermediate poses, which appear between the existing video frames. It allows the synthesized and current poses to form richer motion patterns. Next, the pose distillation divides the body parts into several groups, where it learns the specific part-wise relationship within each group. It degrades the complexity of learning useful part-wise relationships from rich motion patterns and extracts more detailed motion information for fine-grained part groups. We extensively evaluate our method on challenging datasets for dynamic pose estimation, achieving state-of-the-art results.**

***Index Terms*—Intermediate pose synthesis, human pose estimation, hierarchical body structure, information theory.**

I. Introduction

DYNAMIC human pose estimation (dynamic HPE) is an essential task that requires accurately localizing and classifying human body parts, and establishing the complete human skeleton for each frame in the video. It has drawn a lot of attention due to its wide applications including action recognition [1], [2], person re-identification [3], [4], and digital human generation [5]. In real-world applications, the dynamic motions of human body parts yield complex visual patterns, which continually change over time. Thus, dynamic HPE heavily relies on understanding temporal relationship between body parts. It allows the relevant body parts with a strong motion correlation to mutually evidence the part co-existence.

Conventionally, dynamic HPE methods [6], [7] borrow the success of image-based HPE, predicting the body parts and obtain the human skeleton in individual frames without modeling the temporal part-wise relationship. The recent methods [8], [9] employ recurrent neural networks (RNNs) and 3D convolution to extract the temporal features of body parts from multiple frames. Typically, the keyframes' features with remarkable motion changes are used to calculate the temporal features. But this kind of pipeline also suffers from the continuity of human motion. For those visual capturing system with low speed, these keyframes generally have a long-time interval in-between, disallowing dynamic HPE to benefit from the essential motions in short-time periods. On the other hand, the intermediate motions may be unavailable in the video. These motions comprise the inherent human poses, which appear in the short-time intervals missed by the camera with a limited frame rate. For those high-speed capturing systems, the occlusion problem in the continuous motion cannot be avoided. The occlusion of the neighboring frames and the wrong positions of predicted parts can cause the fault trajectory and the misunderstanding of the HPE model when detecting the parts of the current frame. Therefore, the reasonable part moving track in continuous motion is currently urgent to obtain for the dynamic HPE task. In addition, current multi-frame methods focus on the temporal visual context, neglecting the inherent body structure in the dynamic motions. Normally, they learn a shared spatial-temporal representation for body parts, which unavoidably brings the information redundancy and deteriorate the modeling performance.

To tackle these difficulties, we propose to directly generate intermediate human poses based on the captured real frames to complete the continuous human motions, providing efficient temporal contexts. Considering the continuity of human motion and inherent static human body structure, we can naively regard the trajectory of body parts in motion as linear in

Manuscript received 6 March 2024; revised 24 September 2024; accepted 6 February 2025. This work was supported in part by The Hong Kong Polytechnic University (PolyU) under Grant P0048387, Grant P0042740, Grant P0044520, Grant P0043906, Grant P0049586, and Grant P0050657, in part by the PolyU Research Institute for Sports Science and Technology under Grant P0044571, and in part by the National Natural Science Foundation of China under Grant 6247072353. *(Renjie Zhang and Di Lin contributed equally to this work.) (Corresponding author: Ping Li.)*

Renjie Zhang, Xin Wang, and George Baciu are with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: renjie.zhang@connect.polyu.hk; xin1025.wang@connect.polyu.hk; csgeorge@polyu.edu.hk).

Di Lin is with the College of Intelligence and Computing, Tianjin University, Tianjin 300072, China (e-mail: ande.lin1988@gmail.com).

Ruonan Liu is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: ruonan.liu@sjtu.edu.cn).

Bin Sheng is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Pazhou Lab, Guangzhou 510335, China (e-mail: philip.chen@ieee.org).

Ping Li is with the Department of Computing, the School of Design, and the Research Institute for Sports Science and Technology, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

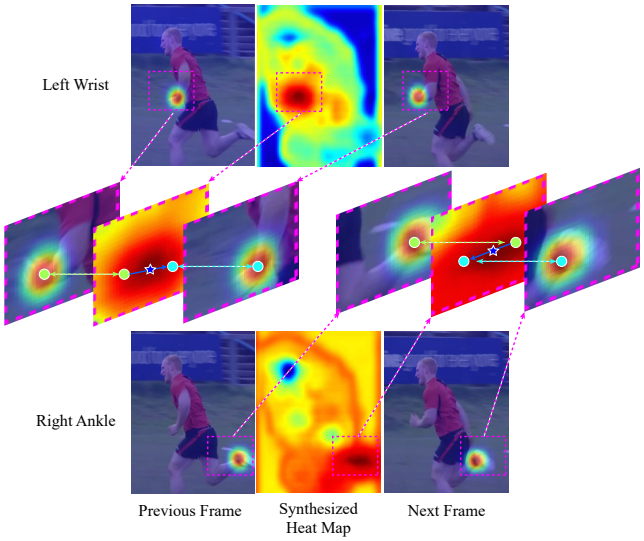


Fig. 1. Temporal-interim poses obtained by our proposed pose synthesizer. For the first and third rows, the first and third columns are adjacent frames, which are attached with the predicted heat maps of the identical part. The second column is heat maps of the same part of the synthesized temporal-interim pose. The pink boxes select the areas with dynamic part locations. In the enlarged boxes in the second rows, we zoom in the parts of the right knee and the left wrist. The green and cyan circles represent the locations of the same part in the adjacent frames. The blue arrows in the middle boxes represent the line trajectory. The blue stars are the synthesized locations of parts in the intermediate interval.

a very short time period. Thus, the generation of temporal interim poses can be simply guaranteed by regression-based supervision. As shown in Fig. 1, our method can synthesize the heat maps of the intermediate pose, indicating the most possible temporal-interim part locations. The existing and intermediate poses together rich complex motion patterns and form the reasonable part trajectory, providing more solid motion information. When there are occlusions or serious prediction of part locations in any of the neighboring frames, the HPE model can rely on the rest of sufficiently enriched reasonable motion contexts. However, they are unsuitable for direct utilization as a hint for predicting the part locations of the existing poses because of the difficulty of achieving useful motion information from these more complex spatial-temporal contexts. To degrade the complexity of information extraction, we propose to learn the temporal contexts of human body parts separately and leverage information-theoretic supervision to extract useful knowledge from the spatial-temporal features of individual body parts. With the specific temporal features of each part group, the model can focus on the fine-grained part motions for each smaller group respectively. The reasonable motion trajectories can be effectively learned in this way and useful information can be obtained via task-relevant objective.

Specifically, we propose a novel framework, called *Temporal-interim Pose Synthesis and Distillation* (see Fig. 2), to harness intermediate motions for dynamic HPE. It consists of two main components: pose synthesis and pose distillation. **(1)** During the processes of image and pose synthesis (see Fig. 2(a)–(b)), we leverage a pretrained Video Frame Interpolation Transformer (VFIT) [10] to generate the visual features representing the intermediate visual status between the existing frames. Along with the learned visual features of adjacent frames, the generated intermediate visual features are given to the pose synthesizer along the temporal dimension. The synthesizer yields the visual features of the interim poses. Moreover, we employ maximum likelihood to make the prediction process is differentiable, building a regression-based objective for the refinement of the interim pose synthesis. It allows the synthesizer to predict the appropriate part locations for the interim poses, which are consistent with the existing poses. **(2)** In the pose distillation (see Fig. 2(c)), which divides the human body parts into multiple groups and focuses on the part-wise relationships within the individual groups. A multi-branch network implements the pose distillation. Each branch has a deformable transformer with multi-head attention, whose training is supervised by a part-related information-theoretic objective. The objective includes a cosine similarity for enchancing the diversity of group features and an mutual information objective for extracting useful information from the refined spatial-temporal features. Our proposed supervision allows each branch to attend to the specific relationship of a group of parts with various locations and extract more detailed motion contexts for fine-grained part groups, thus improving the performance of dynamic HPE.

We evaluate our method on the current benchmark public datasets for dynamic HPE, achieving state-of-the-art results on PoseTrack2017 [11], PoseTrack2018 [12], PoseTrack2021 [13] and sub-JHMDB [14]. And extensive ablation studies are also conducted, validating the effectiveness of each network component and objective we proposed in the framework. Our work makes the following three main contributions:

- We advocate a novel idea of generating temporal-interim human poses in long-time intervals based on the existing captured frames, forming a more continuous motion sequence. With the assumption of a linear trajectory of body parts in motion, we propose a regression-based objective based on the integral to guarantee the consistency of the generated interim poses. The spatial-temporal relationship between body parts can be enriched with the completed motion sequence, bringing improvement to the dynamic human pose estimation.
- We propose a multi-branch network architecture to learn the spatial-temporal contexts of separate body part groups. We further propose information-theoretic objectives to diversify the specific features of different part groups and distill useful information from the extracted features of multiple groups. With feature-level supervision, the specific relationship among groups of body parts can be fully mined from the complex dynamic motion pattern, bringing significant improvements to the dynamic HPE performance.
- Extensive experiments of our model are conducted based on the public dynamic HPE datasets. The state-of-the-art results validate that our proposed framework can synthesize actual interim poses and extract valuable knowledge from them, handling complex cases. Besides, the efficiency experiments demonstrate that our methods achieve higher performance with little computational cost.

II. Related Work

We first introduce traditional image-based human pose estimation, then provide information about the related works on video-based human pose estimation and part-wise relationship learning. These works are relevant to our method, in terms of using the deep network to learn the visual features of the human body parts.

A. Image-Based Human Pose Estimation

Traditional approaches for Image-based human pose estimation are based on the pictorial structure models [15], where the tree-structured graph is used to represent the human body. They rely on prior knowledge of the body structure, thus increasing the complexity of learning the features of the human body. Due to the development of deep learning, most of the recent methods [6], [7], [16], [17] utilize CNNs to extract the visual context from images. The heatmap-based methods [6], [7], [17] have dominated this area. They utilize the likelihood heat maps to represent the locations of the body parts. Some of them [6], [7] directly design different CNN architectures to improve estimation performance of the single-person heat map. Many works have extended this idea and proposed the top-down multi-person pose estimation frameworks, which detect the bounding boxes of humans and use the single-person pose estimator to compute the heat maps for each bounding box. There are also bottom-up methods [17], [18] that compute the heat maps for all body parts. The parts are grouped to form different human poses. Compared with the top-down manner that achieves better accuracy, the bottom-up methods have a faster speed of HPE. However, these methods use directly use the point coordinates with maximum value in the heat map as part locations. The arg-max operation for selecting the coordinates breaks the backward propagation for training the network. In this case, the regression-based methods [19] are considered. IntegralPose [19] has an end-to-end regression framework with the soft-argmax operation to retrieve part locations from heat maps in a differentiable manner. Image-based approaches are only based on static images with spatial features. Thus they perform unsatisfactorily, given videos with motion blur and out-of-focus. Our work utilizes image-based networks as our backbone to provide discrete spatial contexts for each frame, and uses additional architectures to extract continuous temporal information by generating short-time poses. We use the differentiable locations to supervise the visual feature synthesis of short-time poses.

B. Dynamic Human Pose Estimation

To capture the temporal context of human body parts, some recent works [11], [20] divide the multi-frame task into two stages: localizing the body parts in single frames and using temporal smoothing techniques to refine results. LSTMs and 3D CNN are used for temporal smoothing [8], [9], yet requiring much computation. To save computation, some methods [21], [22] use the warping mechanism, which refines the part localization based on sparse keyframes with part-wise annotations. For example, PoseWarper [21] detects poses in the adjacent frames. DCPose [22] and FAMI-Pose [23] introduce a dual temporal direction framework and a feature alignment framework, respectively, for better harnessing the sparse part annotations in keyframes. TDMI [24] exploits dynamic contexts through temporal difference learning and useful information disentanglement. Though keyframes accelerate the speed of dynamic HPE, keyframes normally offer long-time pose information, however missing short-time poses. In contrast to the existing methods, we propose a novel idea to synthesize the short-time poses that are unavailable in the video. Our method lets synthesized and existing poses form more complete motions, eventually benefiting dynamic HPE.

C. Part-wise Relationship Learning

Many methods [25], [26] of HPE respect the characteristics of the human body for modeling the part-wise relationship. These approaches [27]–[30] generally use deep networks to learn the representation of body parts, whose relationships can be adaptively adjusted. PBN [28] regards the pose estimation as homogenous multi-task learning. It divides the human body into multiple groups and uses a individual network module to learn specific features of each group. It learns the relationship between parts within the same group but without interaction between groups. Another array of methods [28], [29] use the tree-like structure to model the part-wise relationships. DLCM [27] introduces a hierarchical compositional model for organizing the body parts. Graph-PCNN [29] introduces a graphical network to extract the relational information of body parts. RPSTN [30] has a pose semantics propagator that transfers the pose semantic information of the current frame to the next frame. In the above methods, the consideration of a complex model of part-wise relationship easily distracts the deep network, which may learn the relationships useless to dynamic HPE. Though the separable part groups restrict the part-wise relationship to the independent groups (or among a few groups), the part-wise relationship learning still needs an effective way to be directly driven by dynamic HPE.

We propose a multi-branch network with multiple transformers for learning part-wise relationships within the separable part groups. We resort to a part-related information-theoretic objective, which drives the part-wise relationship learning to improve the performance of dynamic HPE.

III. Overview

Given a human action video, to estimate the body poses for each frame, we first detect the human in each frame and crop it to obtain an image sequence that contains more than three adjacent human images. Then to leverage richer motion information, we propose a component called pose synthesizer, which takes advantage of the feature maps of adjacent frames to generate feature maps of the intermediate status. We input each neighboring pair of them into a pre-trained VFIT, obtaining rough intermediate visual RGB images for the interval (see Fig. 2(a)). All these feature maps are input into a backbone to compute the visual feature maps, respectively. In the pose synthesizer, we use a deformable transformer to synthesize the intermediate poses between the

TABLE I
IMPORTANT NOTATIONS USED IN THIS PAPER.

$\mathcal{X}$: Function	
$\mathcal{I}$	Integral
$\mathcal{C}$	convolution
$\mathcal{T}$	DETR encoder
$\mathcal{L}/\mathcal{H}/\mathcal{S}$	information/heat map/coordinate loss function
$\mathbb{X}$: Set	
$\mathbb{W}$	temporal window
$\mathbb{I}$	RGB image sequence
$\mathbb{F}$	feature map sequence
$\mathbb{G}$	body part groups
$\mathbb{Z}$	the set of group features
$\mathbb{R}$	the set of real numbers
$\mathbf{X}$ and $\mathbf{x}$: Multi-dimension Representation	
$\mathbf{I}$	RGB image of the human pose
$\mathbf{F}/\mathbf{z}$	visual feature map/group feature map
$\mathbf{H}/\mathbf{P}$	predicted part location heat map/probability heat map
$\mathbf{R}/\mathbf{q}$	2D coordinates of parts/location
X: Constant	
$H \times W$	spatial resolution of the image
$J/N/C$	number of body parts/groups/feature map channels
x: Index and hyperparameter	
t	index of the frame in the sequence
w	index of the frame in the temporal window
δ	temporal bias
k	index of the joints
i, j	index of the part group
γ	hyperparameter of the loss terms
α, β	hyperparameter of the synthesized probability heat map

preliminary frames captured by the camera and the roughly synthesized intermediate frames (see Fig. 2(b)). We employ the transformer to estimate probability heat maps for intermediate poses based on the features of preliminary frames and synthesized frames, obtaining a feature map sequence with continuous motion features. The coordinates of part locations are computed via the integral of probability maps, making the computation of the movement of part locations differentiable. We regard the trajectory of part locations in the short-time period as linear. Then the pose synthesis can be supervised by locating the part locations of the intermediate status into the linear trajectory formed by the poses of adjacent captured frames. The synthesis process is shown in Fig. 3. Besides the inherent synthesis ability of VFIT, our proposed constraint further guarantees the effectiveness of the synthesized poses.

As shown in Fig. 2(c), with the continuous pose feature map sequence, we propose another component called pose distillation. For the input of this component, we combine the feature maps along with the temporal dimension to get a representation which contains the shared temporal and spatial information of the existing and synthesized poses. Then, we divide the human body into multiple groups and adopt a multi-branch network architecture to learn individual features for each part group, computing part-wise group feature maps. Applying convolutions to these group feature maps, the heat maps of part locations can be obtained. By concatenating these group feature maps and corresponding heat maps together, we get the feature maps containing the information for all body parts. Finally, as shown in Fig. 2(d), we combine the obtained feature maps and the preliminary feature maps for the current frame together, applying convolution to the combined feature maps to obtain the heat maps for the overall estimation of the human pose. We further leverage a cosine similarity to diverse the contexts in each feature map and a mutual information objective to distill useful information from different part-group features.

IV. TEMPORAL-INTERIM POSE SYNTHESIS AND DISTILLATION

In this section, details of Temporal-interim Pose Synthesis and Distillation are given. We first give some preliminaries of the framework to describe the preparation and pre-processiing for the input. Then we provide the detailed descriptions of pose synthesis including the network architecture and the supervision. Finally, we present the details of pose distillation. Table I list the definitions of the characters used in the paper.

A. Preliminaries

As illustrated in Fig. 2, we aim to estimate the human pose in the t^{th} frame $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$, where $H \times W$ is the resolution of the input image. We crop the detected human from the frames within a temporal window $\mathbb{W}$. By cropping the human regions, we achieve the image sequence $\mathbb{I} = \{\mathbf{I}_{t+\delta} \mid \delta \in \mathbb{W}\}$. In Fig. 2, we set the length of the window as 5, i.e. $\mathbb{W} = \{-2, -1, 0, 1, 2\}$. By inputting them into a pre-trained VFIT, we obtain the rough RGB image sequence $\mathbb{I}_s = \{\mathbf{I}_{t+\delta,t+\delta+1} \mid \delta \in \mathbb{W}'\}$ for the time interval between each pair of neighboring frames. Here we define $\mathbb{W}' = \{-2, -1, 0, 1\}$. We feed $\mathbf{I}$ and $\mathbf{I}_s$ into a backbone network, producing the preliminary pose feature maps $\mathbb{F} = \{\mathbf{F}_{t+\delta} \mid \delta \in \mathbb{W}\}$ and $\mathbb{F}_s = \{\mathbf{F}_{t+\delta,t+\delta+1} \mid \delta \in \mathbb{W}'\}$ that respectively represent the existing poses and the roughly synthesized interim poses, as shown in Fig. 3. For the simplicity of clarification, we use w^{th} frame to represent any frame in the cropped window of $\mathbb{W}$, i.e., $(w - t) \in \mathbb{W}$.

B. Pose Synthesis

To generate actual interim poses, we build a component called pose synthesizer. As illustrated in Fig. 3, the pair of preliminary pose feature maps $(\mathbf{F}_w, \mathbf{F}_{w+1})$ is passed into the synthesizer, which computes the feature map $\mathbf{F}'_{w,w+1} \in \mathbb{R}^{H\times W\times C}$ of intermeidate pose between the w^{th} and $(w+1)^{th}$

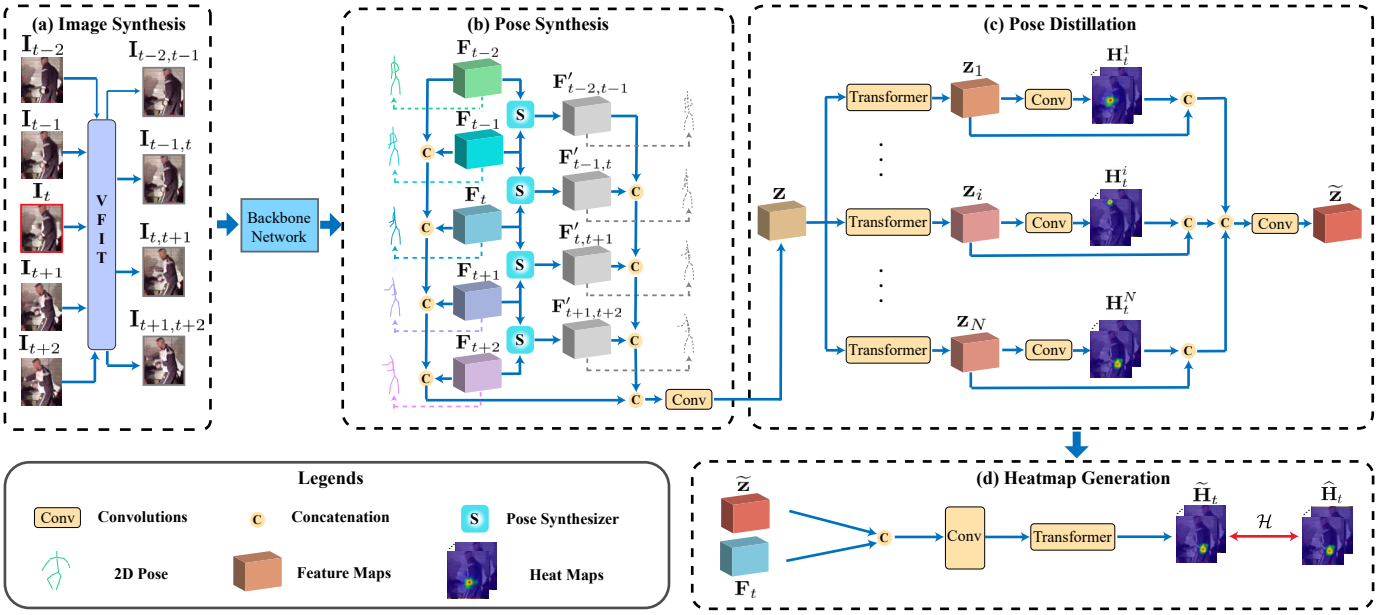


Fig. 2. Illustration of estimation the human pose in the frame $\mathbf{I}_t$. (a) We feed the each pair of neighboring frames (e.g., $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$) into a VFIT for producing the rough intermediate frame. (b) Then we feed the preliminary feature maps (e.g., $\mathbf{I}_t$ and $\mathbf{I}_{t+1}$) of each pair of given frames into the pose synthesize for producing the feature map (e.g., $\mathbf{I}'_{t,t+1}$) of the synthesized intermediate pose. The learning of synthesized pose feature maps is supervised by the loss $\mathcal{L}_s$. (c) We concatenate the preliminary and synthesized feature maps. And we apply convolutions to the concatenation, forming the representation z, which is passed to individual transformers to learn the feature maps $\mathbf{I}_1, ..., \mathbf{I}_N$ of N body part groups. $\mathbf{I}_1$ to $\mathbf{I}_N$ are used to compute the heat maps of part locations in different groups. And we concatenate z_1 to z_N to and use convolutions to get the feature map $\widetilde{\mathbf{z}}$ containing rich useful part-wise information. (d) Finally, we combine $\widetilde{\mathbf{z}}$ and $\mathbf{F}_t$ and apply transformer for predicting the overall heat maps $\widetilde{\mathbf{H}}_t$, supervised by the loss $\mathcal{H}$.

frames. We utilize the visual feature map $\mathbf{F}_{w,w+1}$ from $\mathbb{F}_s$ in the synthesizer to produce rough features for the interim pose. The details are described in Section IV-B. The pose synthesizer computes the final interim pose feature maps in the set $\mathbb{F}' = \{\mathbf{F}'_{w,w+1}\}$. We compute the 2D coordinates of parts $\mathbf{R}_w, \mathbf{R}_{w,w+1}, \mathbf{R}_{w+1} \in \mathbb{R}^{J\times 2}$, based on $\mathbf{F}_w, \mathbf{F}_{w+1}, \mathbf{F}'_{w,w+1}$. J is the number of body parts.

a) Pose Synthesizer: The pose synthesizer leverages adjacent frames $\mathbf{I}_w$ and $\mathbf{I}_{w+1}$ to compute the short-time pose feature map $\mathbf{F}'_{w,w+1}$. We first leverage a VFIT which is pretrained on other dataset to generate the intermediate frame $\mathbf{I}_{t,t+1}$ roughly via interpolation, producing additional knowledge for image feature synthesis. Then we input the existing frames and the synthesized frames into the backbone to obtain preliminary pose feature maps $\mathbf{F}_w$, $\mathbf{F}_{w+1}$ and $\mathbf{F}_{w,w+1}$ of the w^{th}, $(w+1)^{th}$ middle frames in-between. We compute the difference of $\mathbf{F}_{w,w+1}$ and the other two feature maps $\mathbf{F}_w$ and $\mathbf{F}_{w+1}$ for capturing the motion in a short-time slot. And we feed the differences concatenated with $\mathbf{F}_{w,w+1}$ into a deformable DETR $\mathcal{T}$ [31]. In this manner, we allow to $\mathcal{T}$ account for the beginning and ending poses along with the intermediate motion in-between, which form the necessary information for determining the pose between the w^{th} and $(w+1)^{th}$ frames. $\mathcal{T}$ has the deformable attention to comprehensively propagate the information of the given poses and motion to the interim pose feature map $\mathbf{F}'_{w,w+1}$ computed as:

$$\mathbf{F}'_{w,w+1} = \mathcal{T}(\mathbf{F}_{w,w+1},\ \mathbf{F}_{w,w+1} - \mathbf{F}_w, \mathbf{F}_{w,w+1} - \mathbf{F}_{w+1}). \quad (1)$$

Given the preliminary and interim pose feature maps $\mathbf{F}_w$, $\mathbf{F}_{w+1}$, $\mathbf{F}'_{w,w+1}$, we employ an integral operation $\mathcal{I}$ [19] to the heat maps and compute the 2D coordinates $\mathbf{R}_w, \mathbf{R}_{w+1}, \mathbf{R}_{w,w+1}$ of part locations as:

$$\begin{aligned} \mathbf{R}_w =& \mathcal{I}(\mathcal{C}(\mathbf{F}_w)), \quad \mathbf{R}_{w+1} = \mathcal{I}(\mathcal{C}(\mathbf{F}_{w+1})), \\ & \mathbf{R}_{w,w+1} = \mathcal{I}(\mathcal{C}(\mathbf{F}'_{w,w+1})). \end{aligned} \quad (2)$$

b) Supervision for Pose Synthesis: We assume that the intermediate motion change between the w^{th} and $(w+1)^{th}$ frames follows a linear pattern. Between adjacent frames, the identical part of the continual poses form a line trajectory. We let each part of the synthesized pose locate at the middle of the line trajectory. Thus, we only synthesize a pose between each pair of adjacent frames for saving computation. Actually, we can specify any point of the linear trajectory for a part of the synthesized pose. Based on the above assumption, the 2D coordinates of part locations are required for the calculation. Traditional heatmap-based methods employ *maximum likelihood* to heat maps to obtain 2D coordinates. But this process is non-differentiable. To solve this problem, following IntegralPose [19], we estimate probability heat maps and integrate them to compute the part coordinates. For the probability heat maps, we denote the probability heat map of the k^{th} part as $\mathbf{P}^k \in \mathbb{R}^{H\times W}$, the locations in $\mathbf{P}^k$ as $\mathbf{q} \in \mathbb{R}^{1\times 1}$. Then the coordinates $\mathbf{R}^k \in \mathbb{R}^2$ of the k^{th} part can be computed as:

$$\mathbf{R}^k = \sum_{\mathbf{q}\in\Omega} \mathbf{q} \cdot \frac{e^{\mathbf{P}^k(\mathbf{q})}}{\sum_{\mathbf{q}'\in\Omega} e^{\mathbf{P}^k(\mathbf{q}')}}, \quad (3)$$

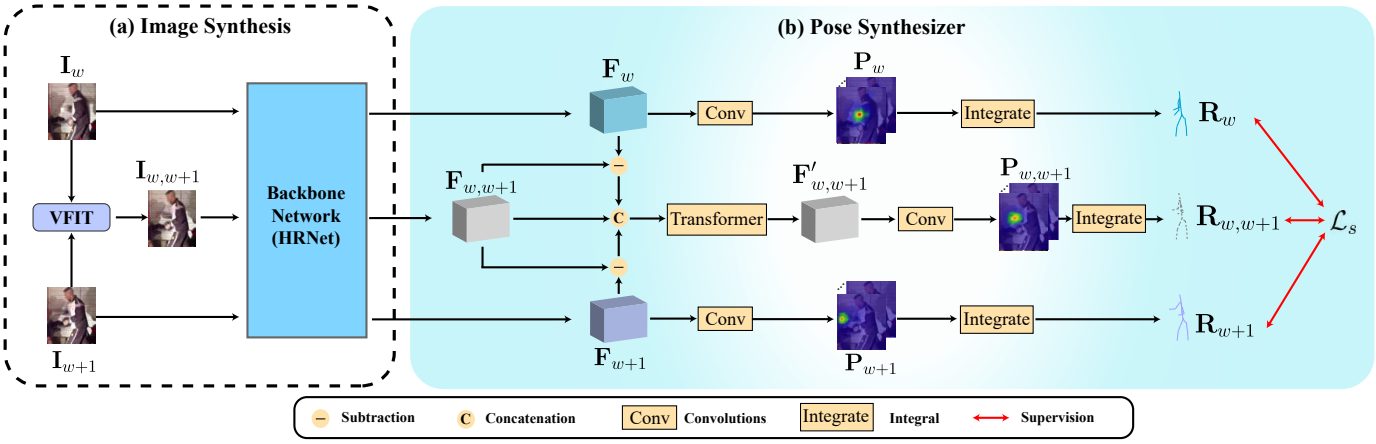


Fig. 3. Illustration of process of interim pose synthesis. (a) First, given neighboring frames $\mathbf{I}_w$ and $\mathbf{I}_{w+1}$, we use a pretrained Video Frame Interpolation Transformer (VFIT) to generate a rough interval image $\mathbf{I}_{w,w+1}$. Then we input all three images into a HPE backbone network to yield corresponding pose features maps $\mathbf{F}_w$, $\mathbf{F}_{w,w+1}$ and $\mathbf{F}_{w+1}$. (b) We feed them into the pose synthesizer, where we combine $\mathbf{F}_{w,w+1}$ and the difference between it and the other two feature maps as the input of a transformer to get a new feature map $\mathbf{F}'_{w,w+1}$. With convolution, the probability heat maps $\mathbf{P}_w$, $\mathbf{P}_{w,w+1}$ and $\mathbf{P}_{w+1}$ are computed. Through integral operations to these probability heat maps, we obtain 2D locations of body parts $\mathbf{R}_w$, $\mathbf{R}_{w,w+1}$ and $\mathbf{R}_{w+1}$. And we leverage a regression-based loss $\mathcal{L}_s$ account for these part locations to supervise the synthesis of interim pose $\mathbf{R}_{w,w+1}$.

where Ω is the domain of $\mathbf{P}^k$ and $\mathbf{q}' \in \mathbb{R}^{1\times1}$ represents all locations in Ω. Concatenating all $\mathbf{R}^k$, we can obtain the 2D coordinates $\mathbf{R}$ of all parts of the overall human body.

Given the part locations $\mathbf{R}_{w,w+1}$ of the synthesized pose, along with locations $\mathbf{R}_w$ and $\mathbf{R}_{w+1}$ of the t^{th} and $(t+1)^{th}$ frames, we compute the loss $\mathcal{S}_w$ as:

$$\mathcal{S}_w = ||\mathbf{R}_w + \mathbf{R}_{w+1} - 2\mathbf{R}_{w,w+1}||. \tag{4}$$

$\mathcal{S}_w$ penalizes any part's location in $\mathbf{R}_{w,w+1}$, which diverges from the trajectory's mid-point determined by the same part in the w^{th} and $(w+1)^{th}$ frames.

Each pair of adjacent frames in the sequence $\mathbb{F}$ and the corresponding roughly synthesized the interim frame from $\mathbb{F}_s$ can be fed into the synthesizer to obtain refined interim poses for computing a regression loss. We compute the overall loss $\mathcal{L}_s$ for the pose synthesis as:

$$\mathcal{L}_s = \sum_{w-t\in\mathcal{W}'} \mathcal{S}_w + 2||\widehat{\mathbf{R}}_t - \mathbf{R}_t||, \tag{5}$$

where $\widehat{\mathbf{R}}_t$ is the ground-truth locations for the t^{th} frame. Here, the second term penalizes the difference between the predicted and ground-truth part locations.

C. Pose Distillation and Heatmap Generation

We combine the feature maps in the sets $\mathbb{F}$ and $\mathbb{F}'$ to compute the representation $\mathbf{z} \in \mathbb{R}^{H\times W\times C}$, which contains the shared information of the existing and synthesized poses. A multi-branch pose distillation learns part-wise relationship from z, as illustrated in Fig. 2(c). Pose distillation divides J body parts into N groups. We denote the groups as $\{\mathbb{G}_i \mid i = 1, ..., N\}$. $\mathbb{G}_i$ contains the part indexes of the i^{th} group. In a branch, a transformer learns the specific part-wise relationship from the group $\mathbb{G}_i$. The transformer takes input as $\mathbf{z}$. Different transformers for the corresponding groups yield the group feature maps in the set $\mathbb{Z} = \{\mathbf{z}_i \in \mathbb{R}^{H\times W\times C} | i = 1, ..., N\}$. Based on $\mathbf{z}_i$, we compute the heat map $\mathbf{H}_t^i \in \mathbb{R}^{H\times W\times||\mathbb{G}_i||}$ for localizing the parts in the group $\mathbb{G}_i$. To let the group feature map $\mathbf{z}_i$ represent the specific relationship within the corresponding group, its heat map $\mathbf{H}_t^i$ is supervised by the heatmap-based loss for the i^{th} group. Furthermore, we employ an information-theoretic loss $\mathcal{L}_{Info}$, which consists of $\mathcal{L}_{var}$, $\mathcal{L}_{com}$, and $\mathcal{L}_{red}$, to distill useful information from complex part-wise relationships. Pose distillation aggregates all $\mathbf{z}_i$ into $\widetilde{\mathbf{z}}$. Then as shown in Fig. 2(d), combing $\widetilde{\mathbf{z}}$ and $\mathbf{F}_t$, we compute the overall heat maps $\widetilde{\mathbf{H}}_t \in \mathbb{R}^{H\times W\times J}$ of all parts. We use an overall heatmap-based loss function $\mathcal{H}$ to supervise $\widetilde{\mathbf{H}}_t$. Below, we give more detailed descriptions of the architecture and supervision of the pose distillation.

a) Multi-Branch Transformers: As illustrated in Fig. 2(c), we concatenate the preliminary and synthesized pose feature maps in $\mathbb{F}$ and $\mathbb{F}'$. And apply a convolution block to it to form a feature map $\mathbf{z}$, which contains more completed motion information. We divide all of the body parts into N groups [28]. To learn part-wise relationships within the individual groups, we build multi-branch transformers, where each branch learns the specific part-wise relationship within a group. For the part group $\mathbb{G}_i$, we use a deformable transformer to extract the i^{th} group feature map $\mathbf{z}^i$ from the shared representation $\mathbf{z}$. We pass $\mathbf{z}^i$ into a convolutional layer to predict the heat map $\mathbf{H}_t^i$, which is concatenated with the group feature map $\mathbf{z}^i$. We pass every pair of concatenated heat map and group feature map into an convolution block, which yields the overall feature map $\widetilde{\mathbf{z}}$ containing information of all parts. As illustrated in Fig. 2(d), in the heat map generation, $\widetilde{\mathbf{z}}$ and the preliminary feature map $\mathbf{F}_t$ are concatenated and input into a transformer block for predicting the final heat maps $\widetilde{\mathbf{H}}_t$ of all parts for the frame $\mathbf{I}_t$.

b) Supervision for Part-Wise Relationship Learning: At first, we compute the losses for supervising the prediction of the overall heat maps $\widetilde{\mathbf{H}}_t$ and part locations in each group at

the same time. The heat map loss can be defined as:

$$\mathcal{H} = ||\widetilde{\mathbf{H}}_t - \widehat{\mathbf{H}}_t|| + \sum_{i=1}^{N} ||\mathbf{H}_t^i - \widehat{\mathbf{H}}_t^i||, \tag{6}$$

where $\widehat{\mathbf{H}}_t$ represents the ground-truth heat map for the whole human body of the frame $\mathbf{I}_t$, the heat map $\mathbf{H}_t^i$ indicates the part localization in the group $\mathbb{G}_i$, and $\widehat{\mathbf{H}}_t^i$ is the ground-truth heat map for the group $\mathbb{G}_i$. The pose distillation learns part-wise relationships from the preliminary and synthesized poses. These relationships should be learned from the groups with diverse visual patterns, thus allowing the relationships to provide richer context information of parts. In addition to the information richness, the learned relationships should be useful for improving the performance of HPE.

To motivate the part-wise relationships to contain rich information, we design the loss $\mathcal{L}_{var}$, which enhances the diversity of different group-wise feature maps in $\mathbb{Z}$. $\mathcal{L}_{var}$ measures the difference between each pair of group-wise feature maps as:

$$\mathcal{L}_{var} = \sum_{1 \le i \le j \le N} \cos[\mathbf{z}_i, \mathbf{z}_j], \quad s.t., \ \mathbf{z}_i, \mathbf{z}_j \in \mathbb{Z}. \tag{7}$$

We denote $cos[\cdot]$ as the cosine similarity. We minimize $\mathcal{L}_{var}$ during the network training.

Furthermore, we compute the losses $\mathcal{L}_{com}$ and $\mathcal{L}_{red}$, letting the group-wise feature maps in $\mathbb{Z}$ contain the task-relevant information. Here, we formulate the lose $\mathcal{L}_{com}$ as:

$$\mathcal{L}_{com} = \mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{F}_t|\widetilde{\mathbf{z}}) - \mathcal{M}(\widehat{\mathbf{H}}_t; \widetilde{\mathbf{z}}|\mathbf{F}_t), \tag{8}$$

where $\mathcal{M}$ is the mutual information between two feature maps. $\mathcal{M}(\widehat{\mathbf{H}}_t; \widetilde{\mathbf{z}}|\mathbf{F}_t)$ represents the amount of task-relevant information contained in $\widetilde{\mathbf{z}}$, which is complementary to the information from preliminary feature maps $\mathbf{F}_t$. The task-relevant information includes part-wise information in $\mathbb{Z}$ and temporal information in $\mathbb{F}$ and $\mathbb{F}'$. We maximize $\mathcal{M}(\widehat{\mathbf{H}}_t; \widetilde{\mathbf{z}}|\mathbf{F}_t)$ to extract additional task-relevant information which includes part-wise information in $\mathbb{Z}$ and temporal information in $\mathbb{F}$ and $\mathbb{F}'$. $\mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{F}_t|\widetilde{\mathbf{z}})$ measures the vanishing of original task-relevant information from $\mathbf{F}_t$. As defined by [24], the first term $\mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{F}_t|\widetilde{\mathbf{z}})$ represents the information in $\mathbf{F}_t$ that is not predictive of $\widetilde{\mathbf{z}}$, so it is the irrelevant information encoded in $\mathbf{F}_t$ regarding $\widetilde{\mathbf{z}}$. Considering that $\widetilde{\mathbf{z}}$ is the learned representation based on $\mathbf{F}_t$, inspired by [24], $\mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{F}_t|\widetilde{\mathbf{z}})$ is actually the vanishing of original task-relevant information from $\mathbf{F}_t$ to $\widetilde{\mathbf{z}}$ during the pose distillation. Minimizing it can prevent excessive information loss in $\mathbf{F}_t$. In other words, we keep the useful information via this. So in the training, we minimize $\mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{F}_t|\widetilde{\mathbf{z}})$ to preserve information of $\mathbf{F}_t$. Overall, $\mathcal{L}_{com}$ is minimized for training.

We use the lose $\mathcal{L}_{red}$ to penalize the redundant information in the group-wise feature maps as:

$$\mathcal{L}_{red} = \sum_{i=1}^{N} \mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{z}_i|\widetilde{\mathbf{z}}), \tag{9}$$

where $\mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{z}_i|\widetilde{\mathbf{z}})$ measures the task-irrelevant information in the group-wise feature map $\mathbf{z}_i$. Here, we minimize $\mathcal{L}_{red}$.

We factorize each term in $\mathcal{L}_{red}$ as:

$$\mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{z}_i|\widetilde{\mathbf{z}}) \to \mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{z}_i) - \mathcal{I}(\mathbf{z}_i; \widetilde{\mathbf{z}}) + \mathcal{M}(\mathbf{z}_i; \widetilde{\mathbf{z}}|\widehat{\mathbf{H}}_t). \tag{10}$$

where the third term $\mathcal{M}(\mathbf{z}_i; \widetilde{\mathbf{z}}|\widehat{\mathbf{H}}_t)$ represents the task-irrelevant information both in $\mathbf{z}_i$ and $\widetilde{\mathbf{z}}$. Naturally, we think that $\mathcal{M}(\mathbf{z}_i; \widetilde{\mathbf{z}}|\widehat{\mathbf{H}}_t)$ will be negligible with sufficient training. So the objective can be simplified as:

$$\mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{z}_i|\widetilde{\mathbf{z}}) \to \mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{z}_i) - \mathcal{M}(\mathbf{z}_i; \widetilde{\mathbf{z}}). \tag{11}$$

Similarly, the two regularization terms in the Eq. (8) can be also simplified as:

$$\begin{aligned} \mathcal{M}(\widehat{\mathbf{H}}_t; \mathbf{F}_t|\widetilde{\mathbf{z}}) &\to \mathcal{I}(\widehat{\mathbf{H}}_t; \mathbf{F}_t) - \mathcal{I}(\mathbf{F}_t; \widetilde{\mathbf{z}}), \\ \mathcal{M}(\widehat{\mathbf{H}}_t; \widetilde{\mathbf{z}}|\mathbf{F}_t) &\to \mathcal{I}(\widehat{\mathbf{H}}_t; \widetilde{\mathbf{z}}) - \mathcal{I}(\widetilde{\mathbf{z}}; \mathbf{F}_t). \end{aligned} \tag{12}$$

Then the minimization of Eqs. (8) and (9), which have the terms of conditional mutual information, can be done by the variational self-distillation [32].

The part related information-theoretic objective can be formulated as:

$$\mathcal{L}_{info} = \gamma_{var}\mathcal{L}_{var} + \gamma_{red}\mathcal{L}_{red} + \gamma_{com}\mathcal{L}_{com}, \tag{13}$$

where the γ_{var}, γ_{red} and γ_{com} are the hyper parameters of the corresponding information-theoretic objectives. Combining all above objectives in the pose distillation and the interim pose synthesis, the overall training objective of the whole framework can be formulated as:

$$\mathcal{L} = \mathcal{H} + \gamma_{info}\mathcal{L}_{info} + \gamma_s\mathcal{L}_s, \tag{14}$$

where γ_{info} and γ_s are the hyper parameters of corresponding loss functions. In the training, we minimize $\mathcal{L}$.

V. Experimental Results

In this section, we show the details of the experimental settings and the results obtained by the proposed model on several video-based human pose estimation datasets. In the tables, the best performance results are bold.

A. Dataset And Evaluation Metric

a) Dataset: We use **PoseTrack2017**, **PoseTrack2018**, **PoseTrack2021** and **Sub-JHMDB** to evaluate our method. PoseTrack2017 is the first large-scale public dataset for multi-person keypoints estimation and tracking in videos. There are 514 videos and 16,219 pose annotations. It has been split into 250, 50, and 214 videos for training, validation, and testing. Then PoseTrack2017 is extened to the PoseTrack2018. There are 1,138 videos with 153,615 posture annotations in the PoseTrack2018. It contains 593/170/375 for training, validation, and testing. PoseTrack2021 is the latest extension of the above two datasets. There are over 420,000 box annotations of humans in this dataset. It provides annotations for more frames in the training and testing sets of PoseTrack2018. In these datsets, each human body is labeled with 15 parts, with extra labels for visibility of parts. a unique person ID and a head bounding box for each person in the annotated frame. In the training set, the 30 frames at the middle of video have annotations. For the validation, every four frames in the

video provide a complete set of part annotations for a human body. Besides, we also conduct experiments on the Sub-JHMDB datasets. It contains 316 videos with 11,200 frames. Only visible parts are annotated. This dataset is divided into three subsets. There are 75% training images and 25% testing images in each subset.

b) Evaluation: The PoseTrack dataset includes three different tasks. Task 1 and Task 2 evaluates the pose estimation performance using the mean average precision (mAP) in a single frame and in videos, respectively. Task 3 focuses on the task of multi-person pose tracking. In this paper, we focus on the task 2, which evaluates the pose estimation performance using the standard evaluation index for human pose estimation, the average precision (AP). We calculate the APs for all body joints of each instance and then average these APs to obtain the mean average precision (mAP).

B. Implementation Details

a) Hyper-Parameters: Our framework is implemented based on PyTorch. We use a single Nvidia GeForce RTX^{TM} 3090 GPU with 24GB memory to train and test the framework. The resolutions of the input image and the feature maps are set to uniform sizes of 384×288 and 96×72. The number of body part groups $\mathcal{N}$ is set to 5, and the division strategy of part groups is the same as PBN [28]. As for the deformable transformer architecture, we follow the original settings in [31]. We set the number of encoders in each branch as 4, the embedding size for attention as 128, and the dimension of the feedforward network as 256. We use an image-based HPE framework called HRNet-W48 [7] as the backbone model in this work. The backbone HRNet-W48 is pre-trained on the COCO dataset. For the hyperparameter , γ_{var}, γ_{red}, γ_{com}, and γ_s are all empirically set as 1, and γ_{info} is set as 0.1. As for the conovlution operations $\mathcal{C}$ in the network, we adopt a set of basic residual convolution blocks with kernel size 3×3. The number of the blocks is set as 3.

b) Bounding Boxes: We follow the two-stage top-down HPE methods [22], [23] to use each bounding box for a single person as the ground truth for training. For validation and testing, we adopt the widely-used person detector [20] to predict the bounding boxes of human. We crop input frames, including the current frame and its temporal neighbors, from the predicted bounding boxes.

c) Training Setting: We set the training epochs to 25. The length of temporal window $\mathcal{W}$ is 4. We utilize Adam solver with an initial learning rate of 1e-4 for optimizing the network. The learning rates are decayed linearly in 10^{th}, 16^{th} and 20^{th} epochs by the factor of 0.1. We conduct data augmentation [23] on the training images by random rotating $[-45°, 45°]$, scaling [0.65, 1.35], and horizontal flipping.

C. Ablation Study

We conduct ablation experiments on the validation set of PoseTrack2017 to evaluate the effectiveness of our method.

TABLE II
ABLATION STUDY ON THE NETWORK COMPONENTS. WE REPORT THE RESULTS ON POSETRACK2017 VALIDATION SET.

Method	Pose Synthesis	Pose Distillation	$\mathcal{S}_w$	mAP
HRNet [7]				77.3
DCPose [22]				82.8
(a)	✓			83.3
(b)	✓		✓	83.8
(c)		✓		85.0
(d)	✓	✓		85.4
(e)	✓	✓	✓	**86.2**

TABLE III
SENSITIVITY TO THE NETWORK ARCHITECTURE. WE REPORT THE RESULTS ON THE POSETRACK2017 DATASET.

Method	Pose Synthesis	Pose Distillation	mAP
HRNet [7]			77.3
(a)	convolution	convolution	84.8
(b)	convolution	transformer	85.8
(c)	transformer	convolution	85.3
(d)	transformer	transformer	**86.2**

a) Effectiveness of Network Components: In Table II, we report the performances of different methods in terms of mAP. These methods are the alternatives without one or more components (i.e., pose synthesis and distillation) of our full model. In the first row of Table II, the backbone HRNet achieves 77.3 mAP. We only use the pose synthesis to compute interim pose feature maps, which are used to compute the shared representation $\mathbf{z}$. We use a simple transformer to predict the human pose based on $\mathbf{z}$. The pose synthesis improves the accuracy by a remarkable margin of 6.5 mAP (see Table II**(b)**), compared to the backbone HRNet. This result demonstrates that the synthesized interim pose features provide more complete motion information. In Table II**(c)**, we only use the pose distillation to learn the part-wise relationship from the preliminary pose feature maps $\mathbb{F}$. Although the synthesized pose feature maps are removed here, the pose distillation still significantly improves the mAP from 82.8 mAP to 85.0 mAP, compared to the backbone HRNet. By combining the pose synthesis and distillation in Table II**(e)** as the full model, we improve the performance to 86.2 mAP, which is better than the result achieved by the pose synthesis only. This result evidences effectiveness of the pose distillation, which learns useful motion information from the mixture of preliminary and synthesized pose feature maps. To illustrate the effectiveness of $\mathcal{S}_w$, in Table II**(a)** and Table II**(d)**, we remove the loss $\mathcal{S}_w$ with different component settings. Compared with DCPose which employs similar temporal information learning module, without $\mathcal{S}_w$, the network in Table II**(a)** cannot regress interim pose, obtaining a lower performance. Similarly, the 85.4 mAP of Table II**(d)** is lower than the 86.2 of the complete framework Table II**(e)**. These results validate the effectiveness of the synthesized interim pose features.

In Table III, we replace the deformable transformer encoders with the deformable convolution blocks, which are used in the pose synthesis and distillation. In Table III**(a)**, we replace

TABLE IV
SENSITIVITY TO THE PROPOSED PART RELATED OBJECTIVES. WE REPORT THE RESULTS ON POSETRACK2017 DATASET.

Loss Functions	$\mathcal{L}_s$	$\mathcal{L}_{var}$	$\mathcal{L}_{com}$	$\mathcal{L}_{red}$	mAP
HRNet [7]					77.3
(a)					83.0
(b)	✓				83.9
(c)	✓	✓			84.1
(d)	✓	✓	✓		85.3
(e)	✓	✓	✓	✓	**86.2**

all transformers with deformable convolutions, yielding 84.8 mAP higher than the backbone HRNet. It means that our network works reasonably without the help of advanced transformers. In Table III**(b–d)**, we add one or more transformers to the pose synthesis and distillation, which produce better results than the model without transformer. It shows a stronger ability of the deformable transformers for learning motion information.

b) Effectiveness of Critical Loss Functions: To investigate the effectiveness of the proposed objectives for dynamic HPE, we change the combination of these objectives and report the mAP in Table IV. In Table IV**(a)**, we only employ the traditional heatmap-based loss function $\mathcal{H}$ to supervise the model. Even without an extra objective, our model (83.0 mAP) still outperforms the baseline HRNet 77.3. In Table IV**(b)**, we add the interim pose synthesis loss $\mathcal{L}_s$ for the pose synthesis, achieving 83.9 mAP. The comparison between Table IV**(a)** and Table IV**(b)** validates the effectiveness of synthesized poses. In Table IV**(c)**, we add the loss function $\mathcal{L}_{var}$ to supervise the part group representations. The improved result of 84.4 mAP shows that diversifying specific features of part groups promotes the extraction of part-wise information. In Table IV**(d)** and Table IV**(e)**, we add $\mathcal{L}_{com}$ and $\mathcal{L}_{red}$ for promoting the relevance between group features. These additional objectives consolidate mAP from 84.1 to 85.3 and 86.2.

c) Sensitivity to Part Grouping: We conduct experiments with different settings of part groups in Table V. The part divisions are obtained by applying spectral clustering to the normalized matrix of mutual information between each pair of body parts [28]. The more considerable number of groups means the more grained division, and each group is smaller. The results of different group settings are shown in Fig. 4. Account for the mean values, it is easy to tell that the performance improves with increasing group number from 1 to 3 or 5. The results demonstrate that smaller groups can bring more detailed part-wise relationships. However, using more groups (i.e., 8 and 15 groups) can hardly result in further improvement. We conjecture that too complex part-wise relationships are unhelpful for the generality of learning motion information. Besides, we can see that for prediction accuracy of head elbow, wrist and ankle obviously changes with the increasing of the group number. On the other hand, the OKS values of the shoulder, hip and knee seems staying static. We conjecture that the more the parts are close to the torso, the parts are more easy to detect, otherwise, the parts are hard to recognize. And the appropriate part grouping setting can help the learning of these parts, such as the setting of 5 groups. In our work, we adopt the division of 5 part groups.

TABLE V
THE GROUP DIVISION SETTINGS FOR THE ABLATION STUDY ON THE GROUP NUMBER.

Group Number	Parts
1	all parts
3	head, upper limb, lower limb
5	head-shoulder, left-lower arm, right-lower arm, thigh, lower limb
8	head-shoulder, left-upper arm, left-lower arm, right-upper arm, right-lower arm, thigh, left-lower leg, right-lower leg
15	right ankle, right knee, right hip, left hip, left knee, left ankle, right wrist, right elbow, right shoulder, left shoulder, left elbow, left wrist, head bottom, nose, head top

TABLE VI
SENSITIVITY TO THE WINDOW LENGTH. WE REPORT THE RESULTS ON POSETRACK2017 SET.

Method	Temporal Window	Speed(FPS)	mAP
FAMI-Pose [23]	$\mathcal{W}=\{-1,0,1\}$	6.9	83.9
FAMI-Pose [23]	$\mathcal{W}=\{-2,-1,0,1,2\}$	3.1	84.8
TDMI [24]	$\mathcal{W}=\{-1,0,1\}$	2.7	84.2
TDMI [24]	$\mathcal{W}=\{-2,-1,0,1,2\}$	1.1	85.7
ours	$\mathcal{W}=\{0\}$	**54.4**	77.3
ours	$\mathcal{W}=\{-1,0,1\}$	2.1	84.8
ours	$\mathcal{W}=\{-2,-1,0,1,2\}$	0.9	**86.2**
ours	$\mathcal{W}=\{-3,-2,-1,0,1,2,3\}$	0.3	85.7
ours	$\mathcal{W}=\{-4,-3,-2,-1,0,1,2,3,4\}$	0.1	85.6

d) Sensitivity to Window Length: In Table VI, we change the length of window that contain adjacent frames and examine the effect on the performance. By enlarging the window from 3 to 5, we improve mAP from 84.8 to 86.2. This is because more adjacent frames provide richer temporal information for synthesizing the interim poses. But using larger windows (i.e., 7 and 9) saturate the performance. We conjecture that the long-term information of more frames is less useful for producing the interim poses and even deteriorates the extraction of temporal-spatial information. Compared with the previous methods FAMI-Pose and TDMI, with same temporal window setting, our method can apparently achieve higher mAP with a competitive speed, respectively.

e) Discussion on Pose Synthesis: We can choose any point of the linear trajectory to regress the part location of synthesized poses. To synthesize different points on the line

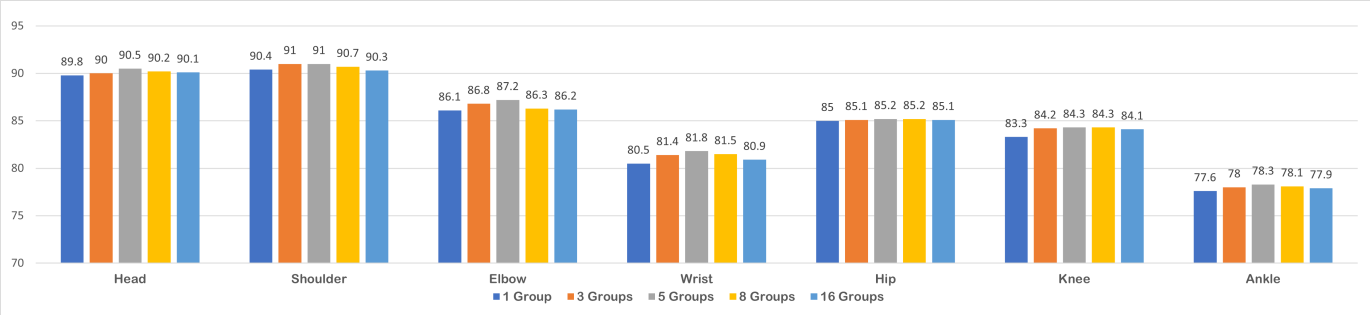


Fig. 4. Sensitivity to the group number. We report the OKS values for different numbers of group in our method on PoseTrack2017 set. We give the OKS values of 7 symmetric parts separately and compare the individual differences among the different group settings.

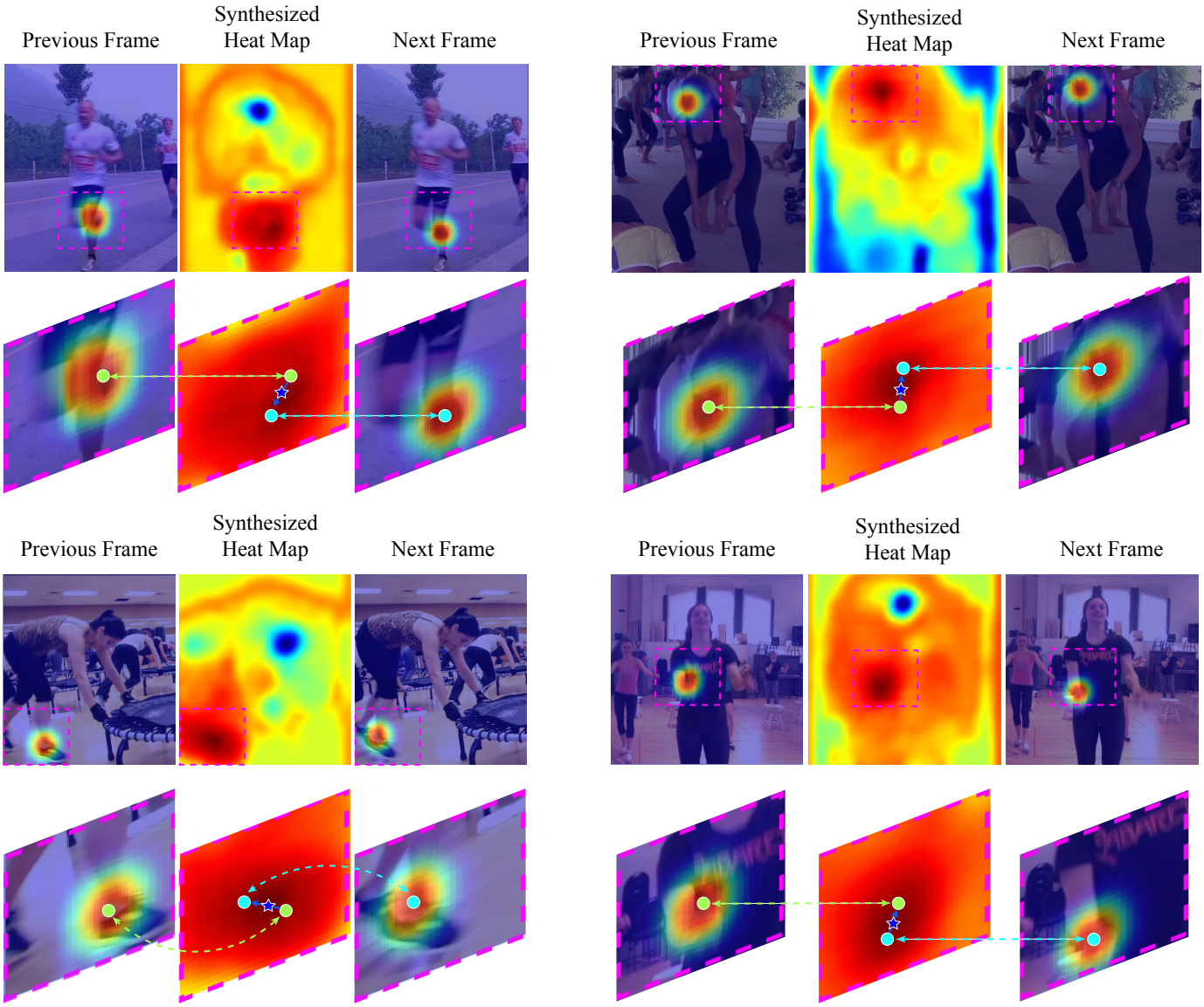


Fig. 5. Visual examples obtained by the pose synthesizer. There are four cases. for each case, The first and third columns are adjacent frames, which are attached with the predicted heat maps of the identical part. the second column is heat maps of the same part of the synthesized interim pose. The pink boxes select the areas with dynamic part locations. In the enlarged boxes in the second and fourth rows, we zoom in the parts of the right knee and the left wrist. The green and cyan circles represent the locations of the same part in the adjacent frames. The blue arrows in the zoomed synthesized map represent the line trajectory. The blue stars are the synthesized locations of parts in the interim interval.

TABLE VII
ABLATION STUDY ON THE POSE SYNTHESIS. WE REPORT THE RESULTS ON POSETRACK2017 DATASET.

Points	$\alpha = 1.5, \beta = 0.5$	$\alpha = 1, \beta = 1$	$\alpha = 0.5, \beta = 1.5$	mAP
(a)	✓			85.1
(b)		✓		86.2
(c)			✓	85.4
(d)	✓	✓		86.0
(e)		✓	✓	86.1
(f)	✓		✓	85.7
(g)	✓	✓	✓	**86.3**

trajectory, we re-formulate the loss $\mathcal{S}_w$ as:

$$\begin{aligned} \mathcal{S}_w &= ||(\alpha \cdot \mathbf{R}_w + \beta \cdot \mathbf{R}_w) - 2\mathbf{R}_{w,w+1}||, \\ s.t.\ \ &\alpha + \beta = 2, \alpha \in [0, 2], \beta \in [0, 2]. \end{aligned} \tag{15}$$

By setting the α and β to different values, the synthesizer can specify different points of the trajectory. The ratio of α/β is lower/higher, the synthesized points are closer to the corresponding part locations in $\mathbf{R}_w$/$\mathbf{R}_{w+1}$. In our work, we set α and β both equal to 1, which means the middle point of the trajectory.

With the above loss, we compare the performances of using three settings of α and β, i.e., ($\alpha = 1.5$, $\beta = 0.5$), ($\alpha = 1$, $\beta = 1$), and ($\alpha = 0.5$, $\beta = 1.5$). We combine these settings to synthesize the interim poses and report the performances in Table VII. In Table VII(a–c), we employ only a single setting of α and β. The comparison demonstrates that the middle point of the line trajectory is more effective. In Table VII(d-g), we combine one or more settings to learn more points, yielding better results than the single setting. More points of the trajectory help the network to recognize the motions and improve the performance. Yet, in consideration of the computational complexity, we use the middle point as default.

f) Discussion on Synthesized Poses: To illustrate the effectiveness of the pose synthesis, we visualize the synthesized poses in Fig. 5. In Fig. 5, we provide 4 example pairs of the adjacent frames from PoseTrack2017, which are attached with the heat maps of the identical part. These heat maps are predicted by the preliminary pose feature maps of the corresponding frames. In the middle column between each pair, we provide the same part's heat map of the synthesized pose, where the part locates at the middle of the linear trajectory formed by the neighboring frames. The crimson areas in the middle column contain the possible locations of the part, showing that the pose synthesis completely understands the intermediate motion. Although our objective is to regress the middle point of the line formed by joint locations from neighboring frames, the learning of the synthesizer focus on a non-linear area in-between, which indicates that the synthesizer acquires motion information from massive training with a large number of images.

D. Comparison with State-of-the-Art Approaches

a) Accuracy Analysis: In Tables VIII and IX, we reproduce different methods and compare their results on PoseTrack2017 and PoseTrack2018 validation dataset. In addition to mAP, we report the average precision (AP) of different parts. Note that other methods work without the help of synthesized poses. Following the TDMI [24], we extend our framework into the multi-stage version (**MS**) where we leverage the multi-stage features extracted by the backbone HRNet to yield preliminary visual feature maps with richer information. Our method outperforms the state-of-the-art methods [24], [41]. The comparison demonstrates that the preliminary and synthesized poses provide more complete motion information. Currently, the comparisons on the test sets of PoseTrack2017 and PoseTrack2018 are unavailable due to the expiration of dataset entrances.

Next, we evaluate our method on the validation set of PoseTrack2021 in Table X and the test set of Sub-JHMDB in Table XI. It should be noted that a large amount of invisible parts exist in the PoseTrack2021. These invisible parts easily lead to erroneous detection of human bodies. Our method still achieves 83.5 mAP, 83.6 mAP and even 84.1 mAP, which are higher than the state-of-the-art methods. Compared to the latest methods, our method also overcomes all state-of-the-art methods on Sub-JHMDB. It shows the robustness of our method. It is noted that with different backbones, our method can achieve better results than the original backbone network. And when compared with the state-of-the-art regression-based method DSTA [41], our method achieves much higher performance on all datasets based on different backbones HRNet and ViTPose respectively. It validates that our method can distillate the useful information from the additional motion contexts by itself instead of heavily relying on the feature extraction ability of any backbone network.

b) Efficiency Analysis: In Table VIII, we also present the parameters, flops and speed of our method, along with all state-of-the-art methods. The speed results are obtained based on a single GeForce RTX™ 3090 GPU card. The flops of our method without/with VFIT [10] are 230G/476G. Compared with the real-time HPE method [37], our methods can achieve much higher prediction accuracy. When it comes to the same type of dynamic HPE methods, although the complexity of our model is a little higher compared with them (183G of FAMI-Pose [23] and 198G of TDMI [24]), we can obtain better results in all datasets, while the speech is competitive. Our full model is only slower than TDMI about 0.2 fps with 0.5 mAP improvement. And when we apply VFIT into the model, we can achieve the state-of-the-art performance in all these datasets. It means that the VFIT can provide extra information for the pose synthesizer, effectively improving the performance. This demonstrates the effectiveness of the usage of VFIT. In addition, our primary contribution is not employing VFIT but generating intermediate motion information with our proposed supervision. As shown in Tables VIII–X, even without VFIT, our model still achieves the best performance with little parameter increase (6.4M) and acceptable Flops increasing (47G) while retaining same inference speed compared with TDMI, which also validates the effectiveness of our whole framework rather than VFIT. And with the huge backbone ViTPose-H, our method can achieve the best performance in all datasets. Using a simple linear regression module for possible intermediate motion trajectory detection

TABLE VIII
COMPARISONS WITH STATE-OF-THE-ART METHODS. WE REPORT THE RECOGNITION ACCURACIES OF DIFFERENT PARTS ON THE POSETRACK2017 DATASET. EACH ACCURACY ACCOUNTS FOR THE SYMMETRIC PARTS.

Method	Backbone	Params	GFLOPs	Speed (FPS)	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
PoseTrack [20]	ResNet-101	-	-	0.8	67.6	70.2	62.0	51.8	60.8	58.8	49.9	60.7
PoseFlow [33]	ResNet-152	-	-	6.7	66.6	73.2	68.2	61.0	67.4	67.0	61.2	66.4
FastPose [34]	ResNet-101	78.68M	154.8	8.7	80.0	80.3	69.5	59.1	71.4	67.5	59.4	70.3
SimeBaseline [6]	ResNet-152	68.6M	35.6	33	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
SE+TE [35]	ResNet-152	-	-	2.3	83.8	81.6	77.1	70.0	77.4	74.5	70.8	77.0
HRNet [7]	HRNet-W48	63.6M	32.9	54.4	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
HRNet+STIP [36]	HRNet-W48	63.6M	35.4	48.1	83.0	82.5	81.6	73.9	76.1	74.9	69.9	77.9
RTMO [37]	CSPDarknet	**44.8M**	**8.1**	**56.1**	81.9	82.3	81.9	72.5	74.1	75.5	68.9	77.1
MDPN [38]	ResNet-152	-	-	-	85.2	88.5	83.9	77.5	79.0	77.0	71.4	80.7
Dynamic-GNN [39]	HRNet-W48	-	-	-	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
PoseWarper [21]	HRNet-W48	71.1M	192.2	4.5	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
DCPose [22]	HRNet-W48	68.0M	46.5	5.0	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
FAMI-Pose [23]	HRNet-W48	64.5M	183	3.1	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
TDMI [24]	HRNet-W48	66.2M	198.6	1.1	90.0	91.1	87.1	81.4	85.2	84.5	78.5	85.7
ViTPose [40]	ViTPose-H	630.7M	121.1	10	88.8	89.6	85.8	80.7	80.4	83.5	75.7	83.8
DSTA-H [41]	HRNet-W48	63.9M	35.7	6.7	89.8	90.8	86.2	79.3	85.2	82.2	75.9	84.6
DSTA-V [41]	ViTPose-H	631.0M	123.9	1.2	89.3	90.6	87.3	82.6	84.5	85.1	77.8	85.6
ours (w/o VFIT)	HRNet-W48	70.8M	230	1.1	89.8	91.0	87.3	82.2	85.3	85.7	78.0	85.9
ours	HRNet-W48	99.1M	476	0.9	91.0	91.2	87.1	82.9	85.1	85.2	78.6	86.2
ours-MS	HRNet-W48	105.3M	492	0.8	**91.1**	**91.5**	87.6	82.1	**85.9**	85.0	79.4	86.4
ours-ViT	ViTPose-H	666.8M	585.0	0.4	90.7	91.1	**87.9**	**83.6**	85.3	**86.3**	**80.0**	**86.7**

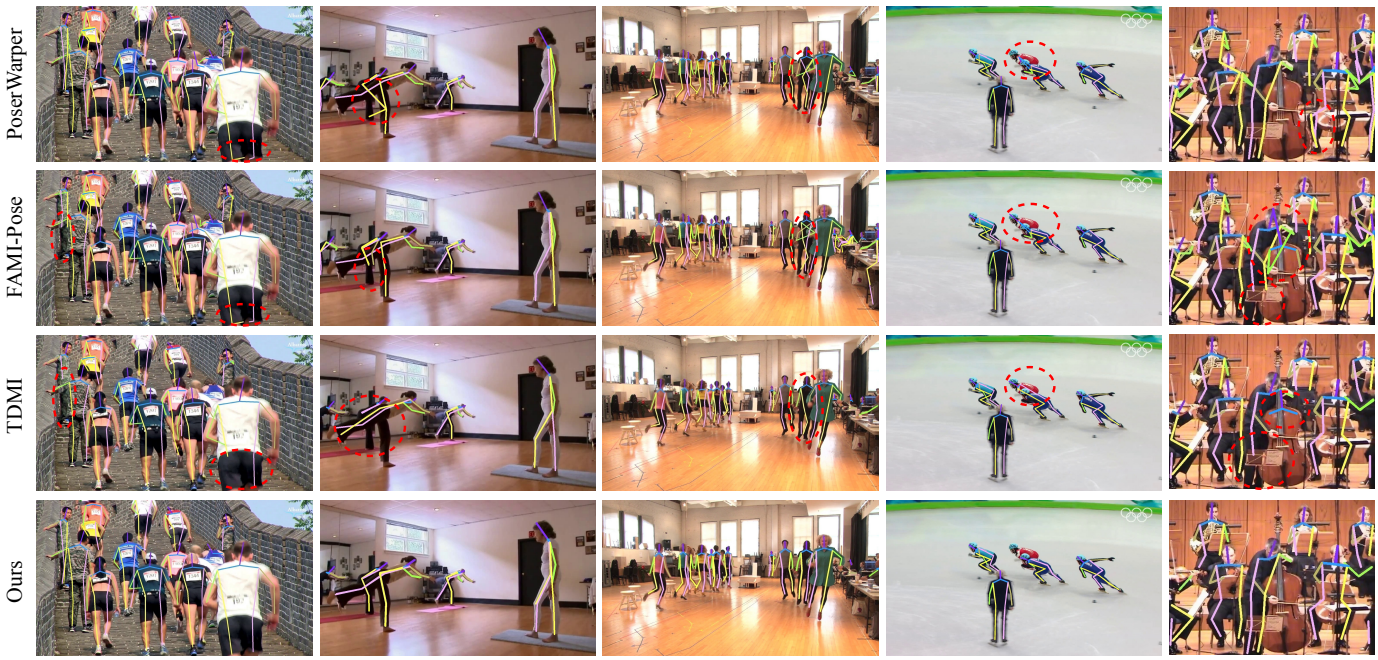


Fig. 6. Part recognition results of different methods. These examples are taken from the PoseTrack2017 and PoseTrack2018 datasets, including different challenging cases, such as complex backgrounds and occlusions. Inaccurate part recognition results are highlighted by the red dotted circles. Compared with the other models, our model can obtain more accurate human pose results in all these hard cases by leveraging more complete motion and part-wise information.

and a multi-branch pose estimation module, our framework can significantly improve dynamic HPE performance based on any backbones in any datasets with a competitive time cost.

c) Visual Results: To illustrate the robustness of our model, we provide some visual results of our model compared with other methods namely PoseWarper [21], FAMI-Pose [23] and TDMI [24] for challenging cases with fast motion, crowded background, or occlusions on PoseTrack17 dataset in Fig. 6. These methods ignore the continuity of human motion and just simply estimate a shared representation, leading to poor performance especially for challenging cases with crowded person instances. In contrast, the accurate human pose estimation results of our methods validate that learning specific information of part-relationships from the completed continuous motion can help to handle visual degradation.

VI. CONCLUSION

We present a novel framework named Intermediate Pose Synthesis and Distillation, which aims to synthesize intermediate human poses between preliminary captured frames, providing more complete motion information for human pose estimation. Our framework utilizes a pre-trained VFIT to generate visual contexts of the short-time interval. And we employ regression-based objective to supervise the intermediate pose synthesis based on the assumption of the linear trajectory in motions. Moreover, we build multiple branches of transformers to distill part-wise relationships, which are learned from the real and synthesized poses. The distillation

TABLE IX
COMPARISONS WITH STATE-OF-THE-ART METHODS. WE REPORT THE RECOGNITION ACCURACIES OF DIFFERENT PARTS ON THE POSETRACK2018 VALIDATION SET. EACH ACCURACY ACCOUNTS FOR THE SYMMETRIC PARTS.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
RMPE [42]	63.9	78.7	77.4	71.0	73.7	73.0	69.7	71.9
MDPN [38]	75.4	81.2	79.0	74.1	72.4	73.0	69.9	75.0
Dynamic-GNN [39]	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
PoseWarper [21]	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
PT-CPN++ [43]	82.4	88.8	86.2	79.4	72.0	80.6	76.2	80.9
DetTrack [9]	84.9	87.4	84.8	79.2	77.6	79.7	75.3	81.5
DCPose [22]	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
FAMI-Pose [23]	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
TDMI [24]	86.2	88.7	85.4	80.6	82.4	82.1	77.5	83.5
DSTA-H [41]	86.2	88.6	84.2	78.5	82.0	79.2	73.7	82.1
DSTA-V [41]	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4
ours (w/o VFIT)	86.4	88.4	85.7	81.2	83.0	82.0	76.9	83.6
ours	86.6	88.9	85.9	80.5	83.1	82.8	77.8	83.9
ours-MS	**87.0**	**89.0**	85.6	81.5	**83.4**	82.4	**78.2**	84.1
ours-ViT	86.9	**89.0**	**86.0**	**81.6**	83.3	**83.3**	78.0	**84.2**

TABLE X
COMPARISONS WITH STATE-OF-THE-ART METHODS. WE REPORT THE RECOGNITION ACCURACIES OF DIFFERENT PARTS ON THE POSETRACK2021 DATASET. EACH ACCURACY ACCOUNTS FOR THE SYMMETRIC PARTS.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
FastPose [34]	56.0	55.9	52.6	48.5	51.0	48.6	44.8	51.4
Tracktor++ w. poses [44]	-	-	-	-	-	-	-	71.4
CorrTrack [45]	-	-	-	-	-	-	-	72.3
CorrTrack w. ReID [45]	-	-	-	-	-	-	-	72.7
Tracktor++ w. corr [44]	-	-	-	-	-	-	-	73.6
DCPose [22]	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose [23]	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
TDMI [24]	85.8	87.5	85.1	81.2	83.5	82.4	77.9	83.5
DSTA-H [41]	87.5	86.6	83.3	78.7	82.7	78.3	73.9	82.0
DSTA-V [41]	87.5	87.0	84.2	81.4	82.3	82.5	77.7	83.5
ours (w/o VFIT)	86.5	86.9	84.8	81.3	83.8	82.3	77.2	83.5
ours	87.1	86.9	84.9	81.0	83.8	82.0	78.0	83.6
ours-MS	87.0	87.6	**85.3**	81.5	**84.0**	82.6	77.9	83.9
ours-ViT	**87.7**	**88.0**	85.0	**81.7**	83.4	**82.8**	**78.3**	**84.1**

TABLE XI
COMPARISONS WITH STATE-OF-THE-ART METHODS. WE REPORT THE ACCURACIES OF DIFFERENT PARTS ON THE SUB-JHMDB DATASET. EACH ACCURACY ACCOUNTS FOR THE SYMMETRIC PARTS.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Avg
Part Models [46]	79.0	60.3	28.7	16.0	74.8	59.2	49.3	52.5
Joint Action [47]	83.3	63.5	33.8	21.6	76.3	62.7	53.1	55.7
Pose Action [48]	90.3	76.9	59.3	55.0	85.9	76.4	73.0	73.8
CPM [16]	98.4	94.7	85.5	81.7	97.9	94.9	90.3	91.9
Thin-slicing Net [49]	97.1	95.7	87.5	81.6	98.0	92.7	89.8	92.1
LSTM PM [8]	98.2	96.5	89.6	86.0	98.7	95.6	90.0	93.6
DKD [50]	98.3	96.6	90.4	87.1	99.1	96.0	92.9	94.0
K-FPN [51]	95.1	96.4	95.3	91.3	96.3	95.6	92.6	94.7
MotionAdaptive [52]	98.2	97.4	91.7	85.2	99.2	96.7	92.2	94.7
FAMI-Pose [23]	99.3	98.6	94.5	91.7	99.2	91.8	95.4	96.0
ViTPose [40]	99.1	98.8	96.8	95.7	99.4	97.5	94.4	97.6
TDMI [40]	99.3	98.9	97.0	93.8	99.3	94.9	95.9	97.1
DSTA-V [41]	99.2	99.0	97.3	94.6	99.1	97.7	95.5	97.6
ours	99.4	99.2	97.3	94.6	99.5	**98.5**	96.7	98.0
ours-ViT	**99.5**	**99.3**	**97.6**	**96.2**	**99.6**	98.3	**96.8**	**98.3**

process offers helpful part-wise relationships for improving the final performance of the dynamic human pose estimation. Substantial experimental results on several evaluation datasets demonstrate the excellent intermediate pose generation performance and strong dynamic feature extraction skill of our proposed framework. However, there is still a limitation in our method: we only adopts specific grouping patterns which are decided previously by manual designs. These fixed patterns unavoidably constraints the part-wise relationship learning. In the future, we will investigate a smarter scheme for extracting part-specific features in motion. We will seek a way to automatically group the body parts during the feature learning process with the visual contexts, instead of using the manually designed grouping patterns. Besides, rather than the current way that extracts part features in the combination of different frames, we will investigate the combination way of part features from individual frames.

REFERENCES

[1] M. Dong and C. Xu, "Skeleton-based human motion prediction with privileged supervision," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 10 419–10 432, 2023.

[2] J. Liu, N. Akhtar, and A. Mian, "Adversarial attack on skeleton-based human action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 4, pp. 1609–1622, 2022.

[3] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4624–4634, 2022.

[4] Z. Wei, X. Yang, N. Wang, and X. Gao, "Flexible body partition-based adversarial learning for visible infrared person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4676–4687, 2022.

[5] D. Liu, L. Wu, F. Zheng, L. Liu, and M. Wang, "Verbal-Person Nets: Pose-guided multi-granularity language-to-person generation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8589–8601, 2023.

[6] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision*, 2018, pp. 472–487.

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

[8] Y. Luo, J. Ren, Z. Wang, W. Sun, J. Pan, J. Liu, J. Pang, and L. Lin, "LSTM pose machines," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5207–5215.

[9] M. Wang, J. Tighe, and D. Modolo, "Combining detection and tracking for human pose estimation in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 085–11 093.

[10] H. Tian, P. Gao, and X. Peng, "Video frame interpolation based on deformable kernel region," in *International Joint Conference on Artificial Intelligence*, 2022, pp. 1349–1355.

[11] U. Iqbal, A. Milan, and J. Gall, "PoseTrack: Joint multi-person pose estimation and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4654–4663.

[12] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "PoseTrack: A benchmark for human pose estimation and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5167–5176.

[13] A. Doering, D. Chen, S. Zhang, B. Schiele, and J. Gall, "PoseTrack21: A dataset for person search, multi-object tracking and multi-person pose tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 931–20 940.

[14] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *IEEE International Conference on Computer Vision*, Dec. 2013, pp. 3192–3199.

[15] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *European Conference on Computer Vision*, 2008, pp. 710–724.

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

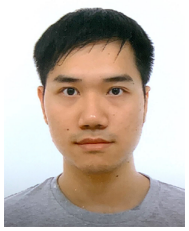
[17] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5385–5394.

[18] Y. Dai, X. Wang, L. Gao, J. Song, F. Zheng, and H. T. Shen, "Overcoming data deficiency for multi-person pose estimation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10 857–10 868, 2024.

[19] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *European Conference on Computer Vision*, 2018, p. 536–553.
[20] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 350–359.
[21] G. Bertasius, C. Feichtenhofer, D. Tran, J. Shi, and L. Torresani, "Learning temporal pose estimation from sparsely-labeled videos," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[22] Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang, and X. Wang, "Deep dual consecutive network for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 525–534.
[23] Z. Liu, R. Feng, H. Chen, S. Wu, Y. Gao, Y. Gao, and X. Wang, "Temporal feature alignment and mutual information maximization for video-based human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 996–11 006.
[24] R. Feng, Y. Gao, X. Ma, T. H. E. Tse, and H. J. Chang, "Mutual information-based temporal difference learning for human pose estimation in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 131–17 141.
[25] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 596–603.
[26] X. Chen and A. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.
[27] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *European Conference on Computer Vision*, 2018, p. 197–214.
[28] W. Tang and Y. Wu, "Does learning specific features for related parts help human pose estimation?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1107–1116.
[29] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-PCNN: Two stage human pose estimation with graph pose refinement," in *European Conference on Computer Vision*. Springer, 2020, pp. 492–508.
[30] Y. Dang, J. Yin, and S. Zhang, "Relation-based associative joint location for human pose estimation in videos," *IEEE Transactions on Image Processing*, vol. 31, pp. 3973–3986, 2022.
[31] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021.
[32] X. Tian, Z. Zhang, S. Lin, Y. Qu, Y. Xie, and L. Ma, "Farewell to mutual information: Variational distillation for cross-modal person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1522–1531.
[33] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," in *British Machine Vision Conference*, 2018.
[34] J. Zhang, Z. Zhu, W. Zou, P. Li, Y. Li, H. Su, and G. Huang, "FastPose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks," *arXiv preprint arXiv:1908.05593*, 2019.
[35] S. Jin, W. Liu, W. Ouyang, and C. Qian, "Multi-person articulated tracking with spatial and temporal embeddings," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5657–5666.
[36] X. Wang, L. Gao, Y. Dai, Y. Zhou, and J. Song, "Semantic-aware transfer with instance-adaptive parsing for crowded scenes pose estimation," in *ACM International Conference on Multimedia*, 2021, pp. 686–694.
[37] P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang, "Rtmo: Towards high-performance one-stage real-time multi-person pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1491–1500.
[38] H. Guo, T. Tang, G. Luo, R. Chen, Y. Lu, and L. Wen, "Multi-domain pose network for multi-person pose estimation and tracking," in *European Conference on Computer Vision Workshops*, 2018, pp. 209–216.
[39] Y. Yang, Z. Ren, H. Li, C. Zhou, X. Wang, and G. Hua, "Learning dynamics via graph neural networks for human pose estimation and tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8070–8080.
[40] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose++: Vision transformer for generic body pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1212–1230, 2024.
[41] J. He and W. Yang, "Video-based human pose regression via decoupled space-time aggregation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1022–1031.
[42] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *IEEE International Conference on Computer Vision*, 2017, pp. 2353–2362.
[43] D. Yu, K. Su, J. Sun, and C. Wang, "Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network," in *European Conference on Computer Vision Workshops*, 2018, pp. 221–226.
[44] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in *IEEE International Conference on Computer Vision*, 2019, pp. 941–951.
[45] U. Rafi, A. Doering, B. Leibe, and J. Gall, "Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos," in *European Conference on Computer Vision*, 2020, pp. 36–52.
[46] D. Park and D. Ramanan, "N-best maximal decoders for part models," in *IEEE International Conference on Computer Vision*, 2011, pp. 2627–2634.
[47] B. X. Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293–1301.
[48] U. Iqbal, M. Garbade, and J. Gall, "Pose for action-action for pose," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017, pp. 438–445.
[49] J. Song, L. Wang, L. Van Gool, and O. Hilliges, "Thin-slicing network: A deep structured model for pose estimation in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5563–5572.
[50] X. Nie, Y. Li, L. Luo, N. Zhang, and J. Feng, "Dynamic kernel distillation for efficient pose estimation in videos," in *IEEE International Conference on Computer Vision*, 2019, pp. 6941–6949.
[51] Y. Zhang, Y. Wang, O. Camps, and M. Sznaier, "Key frame proposal network for efficient pose estimation in videos," in *European Conference on Computer Vision*, 2020, pp. 609–625.
[52] Z. Fan, J. Liu, and Y. Wang, "Motion adaptive pose estimation from compressed videos," in *IEEE International Conference on Computer Vision*, 2021, pp. 11 699–11 708.

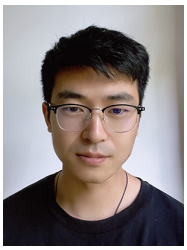
Renjie Zhang received the B.Eng. degree in software engineering from the Sun Yat-sen University, Guangzhou, China, in 2019.

He is currently pursuing the Ph.D. degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include human pose estimation, neural architecture search, deep learning, and 3D human reconstruction.



Di Lin (Member, IEEE) received the B.Eng. degree in software engineering from the Sun Yat-sen University, Guangzhou, China, in 2012, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2016.

He is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. He serves as the senior program committee of IJCAI and associate editor of The Visual Computer. He has been recognized as one of the Stanford/Elsevier World's Top 2% Scientists from 2022 to now. His current research interests include computer vision and machine learning.



Xin Wang received the B.Eng. degree in computer science and technology from the Dalian University of Technology, Dalian, China, in 2017.

He is currently pursuing the Ph.D. degree in computing with The Hong Kong Polytechnic University, Hong Kong. His current research interests include image synthesis, deep learning, computer vision, and diffusion models.

Ruonan Liu (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the Xi'an Jiaotong University, Xi'an, China, in 2013, 2015, and 2019, respectively.

She is currently an Associate Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. She is also an Alexander von Humboldt Fellow with the University of Duisburg-Essen, Duisburg, Germany. She was a Postdoctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, in 2019. She was the recipient of the 2021 Outstanding Paper Award by IEEE Transactions on Industrial Informatics, recognized as one of the World's Top 2% Scientists by Stanford University consecutively from 2021 to now and selected in the Young Elite Scientist Sponsorship Program by CAST in 2022. She is an Associate Editor or Leading Guest Editor of IEEE Transactions on Industrial Cyber-Physical Systems, and Sustainable Energy Technologies and Assessments. Her research interests include machine learning, intelligent manufacturing, and intelligent unmanned systems.

Bin Sheng (Member, IEEE) received the B.A. degree in English and the B.Eng. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2004, and the M.Sc. degree in software engineering from the University of Macau, Macau, in 2007, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2011.

He is currently a Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology. His current research interests include virtual reality and computer graphics.

George Baciu (Senior Member, IEEE) received the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1992.

He was a Professor with the Department of Computing (COMP), The Hong Kong Polytechnic University (PolyU), Hong Kong. He was also a Member of the Waterloo Computer Graphics Laboratory and the Pattern Analysis and Machine Intelligence Laboratory. He is the founding Director of the GAME Laboratory, The Hong Kong University of Science and Technology, Hong Kong, in 1993, and the Graphics and Multimedia Applications Laboratory, COMP, PolyU, in 2000. He has authored or coauthored extensively in computer graphics, image processing, and VR journals and conferences, and was as Chair of many international conferences. His current research interests include information visualization, cognitive computing, virtual reality and computer graphics, with applications to cognitive digital agents, digital twins, motion synthesis, animation, collision detection, geometric modeling, and image analysis.

C. L. Philip Chen (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988.

He is currently a Chair Professor and the Dean of the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Being a Program Evaluator of the Accreditation Board of Engineering and Technology Education (ABET) in the U.S., for computer engineering, electrical engineering, and software engineering programs, he successfully architects the University of Macau's Engineering and Computer Science programs receiving accreditations from Washington/Seoul Accord through Hong Kong Institute of Engineers (HKIE), of which is considered as his utmost contribution in engineering/computer science education for Macau as the former Dean of the Faculty of Science and Technology.

His current research interests include systems, cybernetics, and computational intelligence. He is a Fellow of IEEE, AAAS, IAPR, CAA, and HKIE; a member of Academia Europaea (AE), European Academy of Sciences and Arts (EASA), and International Academy of Systems and Cybernetics Science (IASCYS). He received the IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learnings. He is also a Highly Cited Researcher by Clarivate Analytics in 2018 and 2019. He was a recipient of the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University, in 1988, after he graduated from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA in 1985. He was the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013, the Editor-in-Chief of the IEEE Transactions on Cybernetics (2020-2021), and the IEEE Transactions on Systems, Man, and Cybernetics: Systems (2014-2019), and currently, an Associate Editor of the IEEE Transactions on Fuzzy Systems. He was the Chair of Technical Committee 9.1 Economic and Business Systems of International Federation of Automatic Control from 2015 to 2017, and currently is a Vice President of Chinese Association of Automation (CAA).

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013.

He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 250 top-tier scholarly research articles, pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, realism in non-photorealistic rendering, computational art, and creative media.