REMOTE: Reinforced Motion Transformation Network for Semi-supervised 2D Pose Estimation in Videos

Xianzheng Ma[1,4*], **Hossein Rahmani**[2], **Zhipeng Fan**[3], **Bin Yang**[1], **Jun Chen**[1], **Jun Liu**[4†]

[1]Wuhan University [2]Lancaster University [3]New York University [4]Singapore University of Technology and Design

Abstract

Existing approaches for 2D pose estimation in videos often require a large number of dense annotations, which are costly and labor intensive to acquire. In this paper, we propose a semi-supervised REinforced MOtion Transformation nEtwork (REMOTE) to leverage a few labeled frames and temporal pose variations in videos, which enables effective learning of 2D pose estimation in sparsely annotated videos. Specifically, we introduce a Motion Transformer (MT) module to perform cross frame reconstruction, aiming to learn motion dynamic knowledge in videos. Besides, a novel reinforcement learning-based Frame Selection Agent (FSA) is designed within our framework, which is able to harness informative frame pairs on the fly to enhance the pose estimator under our cross reconstruction mechanism. We conduct extensive experiments that show the efficacy of our proposed REMOTE framework.

Introduction

Human pose estimation, which aims to estimate human body joint locations in image coordinates, is a challenging yet fast-growing research area with wide applications in various fields. Nowadays, the explosive growth of online videos, driven by the ubiquity of mobile devices and sharing activities on social media, demands effective and efficient pose estimation approaches (Nie et al. 2019; Zhang et al. 2020) that can make use of information in videos in a data-driven manner. However, collecting a large scale video dataset with per-frame human pose annotations is often labor intensive and time consuming, limiting the practical scale of such datasets and the development of deep learning-based models.

Instead of collecting fully-labeled videos, a possible remedy would be to craft algorithms to learn from limited labeled data, or more practically, sparsely labeled videos, where temporal information could be additionally leveraged to pass the supervision signal to the unlabeled frames for training the pose estimation networks. PoseWarper (Bertasius et al. 2019) was proposed to propagate pose information from labeled frames to neighboring unlabeled frames, which greatly boosted the pose estimation performance. However, the information is mainly based on local context without considering global context in videos. More recently, the work of (Zhang et al. 2020) proposed a key-frame identification module and a dynamic dictionary to first infer poses for a subset of frames, and then interpolate the key poses to the entire video sequence. However, this method could suffer when the pose sequence to be interpolated becomes complex as the pose dynamics-based dictionary formulation will become challenging.

To effectively exploit the labeled frames and temporal dynamic information in the sparsely annotated videos, in this paper, we propose a novel semi-supervised **RE**inforced **MO**tion **T**ransformation n**E**twork (REMOTE), which consists of three major modules: a Reinforcement Learning (RL)-based *Frame Selection Agent* (FSA), a *Pose Estimator* (PE), and a *Motion Transformer* (MT). Specifically, we design MT to conduct cross frame reconstruction based on paired labeled and unlabeled frames from a video, in which poses estimated by PE are used to guide the frame reconstruction process. Hence, MT learns to warp frames based on the motion dynamics between the current pose and the target pose, which thus enables PE to learn from both labeled frames and unlabeled frames. As a result, under the direct supervision of the labeled frames and extra supervision from the video motion dynamics within the reconstruction process, the capability of PE is enhanced.

However, to enable MT to work effectively, the selected two frames for performing cross reconstruction need to contain moderate variances. This is because if the two frames are temporally too close in the video, the cross frame reconstruction will become trivial and thus MT will not be able to learn motion dynamics sufficiently. In contrast, if the two frames have overly large motion offsets, it may become infeasible for MT to perform effective reconstruction for supervising PE. Therefore, we design an RL-based agent (FSA), guided by a specific reward set, to select informative frame pairs for MT, in order to push our framework towards exploring effective motion dynamics and continuously strengthening the pose estimation performance.

Our main contributions are summarized as follows: **(1)** A novel REinforced MOtion Transformation nEtwork (REMOTE) is proposed to effectively exploit information from both labeled frames and temporal variations in sparsely labeled videos, resulting in a robust and accurate Pose Estima-

*Work done as research intern in SUTD.

†Corresponding Author.

tor (PE). **(2)** A Motion Transformer (MT) module is introduced to learn motion dynamic knowledge in video frames, which is imparted to PE. **(3)** A novel Frame Selection Agent (FSA) driven by a reward set, is proposed to select informative frame pairs, effectively improving the capability of MT and enabling continuous improvement of PE. **(4)** Experimental results show that the proposed REMOTE framework trained on sparsely labeled videos even outperforms many existing models trained on densely labeled videos.

Related Work

Pose estimation. Early works on image-based 2D pose estimation used pictorial structures (Andriluka, Roth, and Schiele 2009; Tian, Zitnick, and Narasimhan 2012; Pishchulin et al. 2013) to represent the keypoints of the human body. Then the work in (Toshev and Szegedy 2014) proposed a deep neural network to directly regress the keypoints, while other works (Chen et al. 2018; Xiao, Wu, and Wei 2018; Jiang et al. 2020) utilized heatmaps to encode the locations of human joints. Besides, HRNet (Cheng et al. 2020; Sun et al. 2019) focused on learning high resolution representations for pose estimation. Apart from image-based pose estimation methods, several other works (Insafutdinov et al. 2017; Girdhar et al. 2018; Yang et al. 2021) addressed 2D human pose estimation from videos. A few works accumulate the temporal information using dense optical flow (Pfister, Charles, and Zisserman 2015; Charles et al. 2016; Song et al. 2017; Jiang et al. 2021) or pose flow (Zhang et al. 2018), while more recently, Recurrent Neural Network (RNN) based methods (Belagiannis and Zisserman 2017; Gkioxari, Toshev, and Jaitly 2016; Lin et al. 2017; Luo et al. 2018) gained more popularity. Another line of work (Wang, Tighe, and Modolo 2020; Liu et al. 2021; Huang et al. 2018; Ruan et al. 2019b,a, 2021) addressed pose estimation and pose tracking simultaneously, to better exploit spatio-temporal information in videos.

Pose estimation in sparsely-labeled videos. Recently, research attention has been drawn to the scenario where only sparsely annotated pose labels are available in videos. DKD (Nie et al. 2019) exploited advances in knowledge distillation to essentially convert the problem of pose estimation to pose matching for fast computation. PoseWarper (Bertasius et al. 2019) introduced a warping-based model to warp the features from the unlabeled frames to the labeled ones, enabling supervised learning on the unlabeled frames using the existing sparse annotations. In the work of (Zhang et al. 2020), a keyframe based computation paradigm was proposed, where the network actively proposes keyframes to query the pose from a pre-trained pose estimator. The estimated poses were then interpolated to obtain the pose in the remaining frames.

Different from the methods in (Bertasius et al. 2019), which is restricted to local propagation of the label information, and (Zhang et al. 2020), which adopts an interpolation module to infer poses, we propose a novel REMOTE framework that takes advantage of an RL-based FSA to *dynamically* select informative frame pairs for the MT, to drive the PE to learn to continuously improve the pose estimation performance based on the sparsely labeled videos.

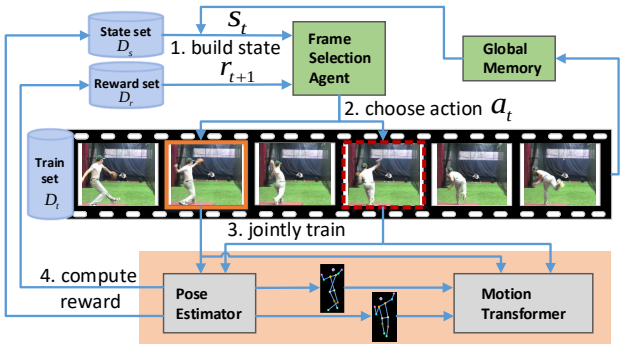

Figure 1: The overview of our REMOTE framework for training 2D pose estimation model from sparsely-labeled videos.

Reinforced Motion Transformation Network

We address the problem of human pose estimation from sparsely labeled videos, where the ground-truth annotations are only provided every K frames. To effectively exploit the information from both the few labeled frames and the temporal variations in videos, we propose a REinforced MOtion Transformation nEtwork (REMOTE), which consists of three modules: a Frame Selection Agent (FSA), a Pose Estimator (PE) and a Motion Transformer (MT). Figure 1 shows the architecture of the proposed REMOTE framework.

In specific, for each labeled frame in the video, FSA learns to select an unlabeled frame for it. Thus an informative pair of labeled and unlabeled frames, containing appropriate motion offset, is obtained. Then the selected informative frame pairs are fed to MT, in order to enable our network to learn from both the information in the labeled frames and the cost-free temporal dynamic information in the video, via a semi-supervised learning manner.

Frame Selection Agent (FSA)

The FSA is designed to select the informative frame pairs to perform cross reconstruction in MT, which is then used to supervise the training of PE. The frame selection can be formulated as a Markov decision process as: $\mathcal{T} = (s_t, a_t, r_{t+1}, s_{t+1})$. Specifically, we model the FSA with a DQN (Mnih et al. 2013), where our FSA scores the combination of the current state s_t and an action a_t, and the unlabeled frame is determined as the state action pair with the highest score. The FSA then receives a reward r_{t+1} as how much improvement the pose estimator E yields by training on this selected frame pair. The FSA then updates its state s_t and moves to the next state s_{t+1}.

As shown in Fig. 1, we split the original training dataset into three disjoint subsets: training set $\mathcal{D}_t$, state set $\mathcal{D}_s$, and reward set $\mathcal{D}_r$. The FSA learns the policy by playing the frame pair selection game on $\mathcal{D}_t$. Meanwhile, both $\mathcal{D}_s$ and $\mathcal{D}_r$ contain only a small subset of labeled frames, and are employed for evaluating the current state (i.e., the performance) of the PE and the performance gain brought by taking a specific action. Below we describe the state, action, and reward formulation in detail.

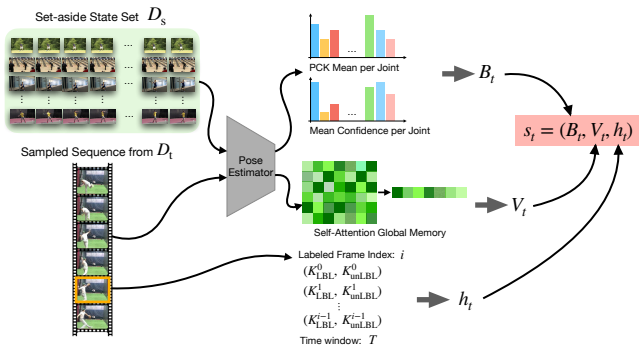

Figure 2: Illustration of state representation computation

State Representation Motivated by (Wu et al. 2019, 2020; Wang et al. 2021), the FSA receives information from the state signal s_t to make the frame selection decision. To select appropriate frame pairs, the information fed to the FSA should contain: 1) current performance of the pose estimator (to see how much room the pose estimator E has for improvement); and 2) global contextual information in the video (to see where the informative frames might be). For the performance information, $\mathcal{D}_s$ is used to evaluate the performance of the pose estimator E. To ensure $\mathcal{D}_s$ is representative enough, we sample $\mathcal{D}_s$ to match the distribution of the original training dataset, and thus the improvement obtained on $\mathcal{D}_s$ could be transferred to $\mathcal{D}_t$ successfully. For the global contextual information, we introduce a global memory module to encode the pose evolution over time, allowing FSA to navigate through the temporal information more efficiently. Besides, the agent needs to know additional information about the previously executed actions and the video sequence to repeat the whole episode of the game. Therefore, as shown in Fig. 2, we represent the state of FSA at the time step t as a tuple $s_t = (B_t, V_t, h_t)$, where B_t, V_t, h_t denote performance information, global contextual information, and additional information, respectively.

For B_t, we adopt the standard PCK evaluation metric to measure the performance of the pose estimator (detailed in Sec 4). The performance is evaluated on the separate state set $\mathcal{D}_s$ to avoid overfitting. Empirically, we notice that the overall PCK score is not representative enough to encode the performance. Therefore, we encode the *PCK score distribution* for each joint as the performance indicator. Besides, the confidence of the pose estimator is also valuable for evaluation. Therefore, we use the average max heatmap response over each joint as another indicator of the performance. For V_t, it is necessary to encode the temporal pose evolution over the video. To this end, we first compute the pose heatmap features frame by frame. Then a self-attention mechanism (Vaswani et al. 2017) is used to weighted combine the heatmap features, forming the final global memory V_t. Finally, h_t serves as auxiliary information, which encodes the index k of the current given labeled frame in the video, the time window T for limiting the action space, and the past selected pairs represented as a dictionary of frame pair's indices.

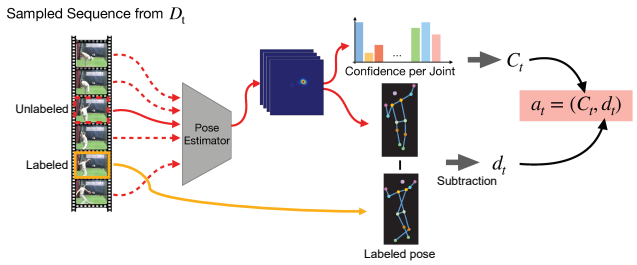

Figure 3: Illustration of action representation computation

Action Representation In our framework, the FSA takes an action by selecting an unlabeled frame from a pool of unlabeled frames to pair with the given labeled frame. Note that the FSA is supposed to select a frame pair with appropriate motion offset. To this end, the action representation a_t should encode the motion offset related information. As shown in Fig. 3, we encode d_t, which is the difference (subtraction) between the estimated pose coordinates from a candidate unlabeled frame and the ground-truth pose coordinates of the labeled frame, as the main element of our action representation a_t. However, since the estimated pose from the unlabeled frame is not always accurate (particularly at the beginning of the training), the calculated difference alone is not good enough for encoding the motion offset. Therefore, we propose to additionally encode the confidence score C_t of the pose estimated from the unlabeled frame, i.e., the max heatmap response distribution over each joint, as another element of a_t. Finally, we represent the action representation as $a_t = (C_t, d_t)$.

To achieve efficient frame selection, we limit each action in a restricted action space $\mathcal{A} = [t_k - T, t_k + T]$, where t_k is the given labeled frame, and T is the restricted time window. Based on the well defined action space, our proposed FSA can not only jump forward to seek future informative frames, but also go back to re-examine past information.

Reward Function After jointly training the PE and the MT with the frame pair selected by FSA, the updated PE will be evaluated on the held-out dataset $\mathcal{D}_r$ at the t^{th} time step to get the PCK scores (PCK_t). The reward signal is then estimated, by checking if incorporating the newly picked frames could boost the pose estimation accuracy. To this end, we define the reward as:

$$r_t = \text{PCK}_t - \max_{e \in (0, t-1]} \{\text{PCK}_e\}. \quad (1)$$

According to Eq (1), the FSA is rewarded only when the newly updated pose estimator outperforms all previous ones. Otherwise, the FSA receives a penalty proportional to the drop in accuracy. Therefore, the reward function acts as an explicit proxy indicating the effectiveness of the selected frames on the performance of the PE.

Optimization As mentioned above, we employ three disjoint splits of the original training dataset ($\mathcal{D}_t$, $\mathcal{D}_s$, $\mathcal{D}_r$) to tailor the FSA for improving PE. The FSA formulated as a DQN (Mnih et al. 2013; Casanova et al. 2020) determines the action based on the Q values of the state-action pair

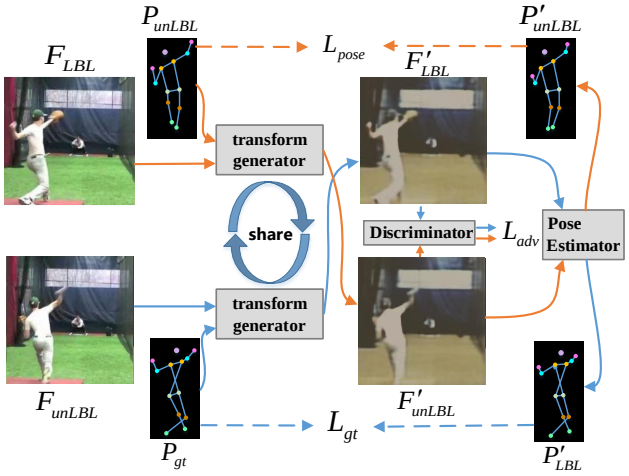

Figure 4: Pipeline of jointly training the Motion Transformer (MT) and the Pose Estimator (PE).

(s_t, a_t). The action corresponding to the maximum Q value is raised, which indicates the index of the selected unlabeled frame to pair with the labeled frame. The proposed DQN agent is trained on the $\mathcal{D}_t$ and the rewards are computed on the held-out split $\mathcal{D}_r$. We train the DQN agent based on the Temporal Difference (TD) error (Sutton 1988), which is computed based on samples from the experience replay buffer $\mathcal{E}$:

$$\mathbb{E}_{\mathcal{T}\sim\mathcal{E}}\left[\left(y_t - Q\left(s_t, a_t; \phi\right)\right)^2\right], \tag{2}$$

where y_t denotes the TD target.

Pose Estimator (PE)

We adopt the Simple Baseline (Xiao, Wu, and Wei 2018) with ResNet50 (He et al. 2016) as our baseline pose estimator E. Given a pair of an annotated frame F_{LBL} and an unlabeled frame F_{unLBL} selected by the FSA, the pose estimator E infers the poses P_{LBL} and P_{unLBL} correspondingly:

$$P_{\mathrm{LBL}} = E(F_{\mathrm{LBL}}); P_{\mathrm{unLBL}} = E(F_{\mathrm{unLBL}}), \tag{3}$$

which are then fed to the MT to exploit the dynamic information in the frame pair. Note that the pose estimator E is of our interest after training.

Motion Transformer (MT)

The MT not only learns from the labeled frames but also exploits temporal dynamic information in the paired frames. As human motions can have different speeds and amplitudes, the motion offset between the two frames could vary. If we can model and learn the variety in the motion offset and pass the knowledge down to the PE, the PE will be robust and more accurate. Motivated by (Gui et al. 2018; Hernandez, Gall, and Moreno-Noguer 2019), we propose to perform cross reconstruction mechanism between the two selected frames and employ the image level and pose-related feature level errors as the supervision signals.

Specifically, the MT module is employed to: 1) reconstruct the labeled frame from its ground-truth pose and the unlabeled frame; and 2) reconstruct the unlabeled frame from its estimated pose and the labeled frame. Note that if this cross frame reconstruction mechanism functions effectively, the estimated pose has to be accurate; Otherwise, the transform generator will fail to accurately manipulate the subject in the input frame. Using the estimated pose as a bridge, it implicitly encourages our pose estimator E to provide more accurate estimations.

We additionally introduce a ConvNet-based discriminator to train the MT in an adversarial manner by enforcing more constraints on the predictions of the MT and consequently the PE. The discriminator estimates if a given pose and a generated person image are a match, based on: 1) the quality of the input image; and 2) if the pose of the subject in the image matches the given pose. The discriminator further promotes the accuracy of our pose estimator E.

Transform Generator Inspired by the recent advances in pose guided person image generation (Ma et al. 2017; Zhu et al. 2019), we employ a similar pose guided transform generator in our MT. The traditional generators in (Ma et al. 2017; Zhu et al. 2019) take a source image containing a subject and a target pose as input, and generate a realistic and appearance consistent image of the source subject in the target pose. In order to simultaneously utilize the generator's ability in modeling motion dynamic knowledge in frame pairs and generating images which are shape-consistent with the target poses, we modify the traditional generators and propose a cross reconstruction mechanism as follows. Given a frame pair (F_{unLBL}, F_{LBL}), in order to perform the cross reconstruction mechanism effectively, we first pass the unlabeled frame F_{unLBL} and the ground-truth pose P_{gt} of the labeled frame F_{LBL} through the transform generator to generate the reconstructed frame F'_{LBL}. Then, the labeled frame F_{LBL} and the estimated pose P_{unLBL} of the unlabeled frame F_{unLBL} are fed to the generator to produce the reconstructed frame F'_{unLBL} as follows:

$$F'_{\mathrm{LBL}} = G(F_{\mathrm{unLBL}}, P_{\mathrm{gt}}) \tag{4}$$

$$F'_{\mathrm{unLBL}} = G(F_{\mathrm{LBL}}, P_{\mathrm{unLBL}}). \tag{5}$$

In addition to the implicit cross reconstruction loss, we could also add an explicit supervision on the reconstructed frames by checking if the pose estimated from the reconstructed frame matches with the one extracted from the actual frame. As shown in Fig. 4, the reconstructed frames F'_{LBL} and F'_{unLBL} are then fed back to the pose estimator E, to generate their corresponding pose heatmaps P'_{LBL} and P'_{unLBL}:

$$P'_{\mathrm{LBL}} = E(F'_{\mathrm{LBL}}); P'_{\mathrm{unLBL}} = E(F'_{\mathrm{unLBL}}) \tag{6}$$

Discriminator To obtain more realistic reconstruction and further encourage the pose transfer generator to faithfully reconstruct frames with poses that match the target poses, a shape discriminator is introduced. Given a frame (F_{i} or F'_{i}) and the estimated pose P_{i}, $\mathrm{i} \in \{\mathrm{LBL}, \mathrm{unLBL}\}$, the discriminator scores four probabilities on the consistency of the pose and the input frame:

$$p_{\mathrm{i}}^{(\prime)} = Discriminator([F_{\mathrm{i}}^{(\prime)}, P_{\mathrm{i}}]), \tag{7}$$

where $[\cdot,\cdot]$ denotes the channel-wise concatenation. p_{i} scores the original input frame F_{i} and pose P_{i} pair, whereas p'_{i} evaluates the reconstructed frame F'_{i} and pose P_{i} pair. By providing both pose and appearance information as the input of the discriminator, we encourage the pose transfer generator to reconstruct more realistic frames while taking care of the pose consistency of the reconstructed frames.

Loss function To successfully train MT, we need comprehensive supervisions at each training stage of our MT. Firstly, we should exploit the ground-truth labels whenever possible. We denote this loss as $\mathcal{L}_{\mathrm{gt}}$, which is the mean square error between the predicted and the ground-truth heatmaps, and is only defined for the labeled frames:

$$\mathcal{L}_{\mathrm{gt}} = \frac{1}{N}\sum_{t=1}^{N}(||P^{\mathrm{t}}_{\mathrm{LBL}} - P^{\mathrm{t}}_{\mathrm{gt}}||_2 + ||P'^{\mathrm{t}}_{\mathrm{LBL}} - P^{\mathrm{t}}_{\mathrm{gt}}||_2), \quad (8)$$

where P_{LBL} and P'_{LBL} are heatmaps estimated from the labeled frame F_{LBL} and its reconstructed version F'_{LBL}, respectively, P_{gt} is the ground-truth heatmap of the frame F_{LBL}, and N is the number of labeled frames.

To encourage vivid reconstruction as well as the consistency between the target poses and the pose in the reconstructed frames, we introduce an adversarial loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{adv}} = &\mathbb{E}_{F_{\mathrm{i}}\in\mathcal{F},P_{\mathrm{i}}\in\mathcal{P}}\left[\log(p_{\mathrm{LBL}}p_{\mathrm{unLBL}})\right] \\ &+ \mathbb{E}_{F'_{\mathrm{i}}\in\hat{\mathcal{F}},P_{\mathrm{i}}\in\mathcal{P}}\left[\log((1-p'_{\mathrm{LBL}})(1-p'_{\mathrm{unLBL}}))\right],\end{aligned} \quad (9)$$

where $\mathrm{i}\in\{\mathrm{LBL},\mathrm{unLBL}\}$, and $\mathcal{P}$, $\mathcal{F}$ and $\hat{\mathcal{F}}$ denote the distribution of human pose heatmaps, real and reconstructed human frames, respectively. We encourage the realistic reconstruction here though it is not our ultimate goal, as it will help estimate the following pose consistency loss.

The $\mathcal{L}_{\mathrm{adv}}$ and $\mathcal{L}_{\mathrm{gt}}$ have no direct restrictions on the poses in the reconstructed frames. To address this issue, we propose the pose consistency loss $\mathcal{L}_{\mathrm{pose}}$ to constrain the poses of the reconstructed frames to be similar to the estimated poses of the input frames on a semantic level.

The pose estimator E is utilized again to extract pose related features from F_{LBL}, F'_{LBL}, F_{unLBL}, and F'_{unLBL} at four intermediate layers of the pose estimator E:

$$\mathcal{L}_{\mathrm{pose}} = \mathbb{E}_{F_{\mathrm{i}}\in\mathcal{F},F'_{\mathrm{i}}\in\hat{\mathcal{F}}}\sum_{k=1}^{4}\sum_{\mathrm{i}}\left\|E_k(F'_{\mathrm{i}}) - E_k(F_{\mathrm{i}})\right\|_2, \quad (10)$$

where $\mathrm{i}\in\{\mathrm{LBL},\mathrm{unLBL}\}$ and E_k corresponds to the k^{th} layer feature map of the pose estimator E.

Finally, our complete objective function is defined as the weighted combination of all the aforementioned losses:

$$\mathcal{L}_{\mathrm{full}} = \lambda_1\mathcal{L}_{\mathrm{gt}} + \lambda_2\mathcal{L}_{\mathrm{adv}} + \lambda_3\mathcal{L}_{\mathrm{pose}}, \quad (11)$$

where $\lambda_i(i\in\{1,2,3\})$ controls the relative importance of the three objectives. Empirically, we set the λ_1,λ_2 and λ_3 in (11) as 1, 0.25, and 0.5. We aim to solve the minimax game:

$$E^*, G^* = \arg\min_{E,G}\max_{D}\mathcal{L}_{\mathrm{full}}. \quad (12)$$

Training and testing

During training, we use the MPII (Andriluka et al. 2014) pre-trained pose estimator and Market-1501 (Zheng et al. 2015) pre-trained transfer generator to initialize our pose estimator and MT, respectively. This provides a relatively good accuracy to start with. The training and testing procedure of our REMOTE framework can be summarized as follows:

1. The state set $\mathcal{D}_s$ is used to compute state s_t for FSA.
2. FSA finds the first labeled anchor frame in the first video sequence in $\mathcal{D}_t$ and generates action a_t, i.e., proposing an unlabeled frame F^t_{unLBL} to pair with this labeled frame F^t_{LBL} with an ϵ-greedy policy.
3. Given the selected pair, the MT performs cross frame reconstruction and passes the extracted motion knowledge to improve the pose estimator E by updating its parameters.
4. The reward r_{t+1} is computed based on the performance improvement of the E on $\mathcal{D}_r$ and then fed to the FSA.
5. The FSA updates its parameters and moves to the next labeled anchor frame.
6. Repeat steps 1 to 5 until the reward becomes negative for five consecutive times. Then, the FSA switches to the next video sequence in $\mathcal{D}_t$. The training stops after processing all the video sequences in $\mathcal{D}_t$.
7. The pose estimator E is evaluated on the test dataset.

Experiments

Datasets. We evaluate our proposed REMOTE framework on two widely used video pose estimation datasets: Penn Action (Zhang, Zhu, and Derpanis 2013) and Sub-JHMDB (Jhuang et al. 2013).

Penn Action (Zhang, Zhu, and Derpanis 2013) is a large-scale unconstrained video dataset containing 2326 video sequences of 15 different actions, with 1258 videos for training and the rest for testing. On average, each video contains 70 frames. We uniformly sample 3 videos from each action category in the original training dataset to build the $\mathcal{D}_s$. For the $\mathcal{D}_r$, we uniformly sample 60 videos (4 videos per action category) from the remaining training dataset. The rest (1153 videos in total) is considered as the $\mathcal{D}_t$.

Sub-JHMDB (Jhuang et al. 2013) Following (Nie et al. 2019; Luo et al. 2018; Zhang et al. 2020), we use the subset of JHMDB (i.e., Sub-JHMDB) to evaluate our method. This subset contains 316 full-body person videos and 12 different action categories. 15 body joints are labeled on each frame excluding invisible joints. Sub-JHMDB has 3 different split schemes with a training and testing ratio of roughly 3:1. We train our model separately and report the average results over three splits. For each split scheme, we uniformly sample 12 videos (1 video per action category) from the corresponding training dataset to build the $\mathcal{D}_s$. 12 videos are sampled from the remaining dataset to build the $\mathcal{D}_r$. The remaining videos are considered as the $\mathcal{D}_t$.

Evaluation Metrics. We measure the performance of the PE with the standard Percentage of Correct Keypoints (PCK) (Yang and Ramanan 2013). A joint is considered to be correct if it lies within a predefined threshold αL, where α is a scaling coefficient and is conventionally set to 0.2 while L is the reference distance, which is set to

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Avg
Baseline, w/o fine-tune	88.5	79.5	72.3	68.9	84.5	78.3	78.0	77.8
Pose Estimator E+$\mathcal{L}_{\text{gt}}$	88.7	79.7	72.6	69.5	84.8	78.7	78.2	78.1
MT+Random	92.2	81.4	73.8	70.5	85.0	79.2	78.9	78.6
MT+Uniform-3	88.9	79.7	73.5	69.7	85.3	78.7	79.1	78.5
MT+Uniform-6	90.3	80.5	74.2	69.8	85.6	78.5	79.3	78.9
MT+Uniform-9	90.1	79.5	73.1	69.4	85.2	77.9	78.8	78.3
MT+FSA - $\mathcal{L}_{\text{gt}}$	88.6	79.9	72.5	69.4	84.9	78.6	78.4	78.2
MT+FSA - $\mathcal{L}_{\text{adv}}$	89.4	80.3	73.2	70.1	85.2	79.7	79.1	78.8
MT+FSA - $\mathcal{L}_{\text{pose}}$	89.0	80.8	74.1	70.4	85.1	80.7	78.8	79.1
MT+FSA	**93.8**	**82.9**	**76.2**	**71.3**	**85.9**	**81.4**	**80.3**	**80.8**

Table 1: Ablation study on the *frame sampling mechanism* and *Motion Transformer (MT) losses* for Sub-JHMDB dataset with an average of 7.2 labeled frames per video. The stricter $\text{PCK}_{\text{torso}}$ is used.

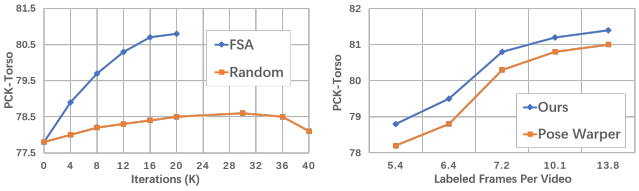

Figure 5: (a) Performance comparison of sampling mechanisms. Our proposed FSA converges after $20K$ iterations, while Random selection converges at $30K$ iterations. (b) Comparison on $\text{PCK}_{\text{torso}}$ w.r.t the number of labeled frames. Our proposed method outperforms PoseWarper (Bertasius et al. 2019) under any annotation density.

$L = max(H, W)$ in the typical PCK_{body} setting (Song et al. 2017; Luo et al. 2018). H and W denote the height and width of the bounding box containing a person instance. Following (Nie et al. 2019; Luo et al. 2018), we additionally adopt the stricter $\text{PCK}_{\text{torso}}$, whose reference distance L is set to the torso diameter (Luo et al. 2018).

Implementation details. To improve the diversity of training data, we perform data augmentation following conventional strategies (Zhang et al. 2020; Bertasius et al. 2019). For the Penn Action, the scaling factor ranges from 0.8 to 1.4 while for Sub-JHMDB it ranges from 1.2 to 1.8. We use GT bounding box to crop each person and pad to a fixed size (384×384). We use ResNet50 as the backbone of (Xiao, Wu, and Wei 2018). Following PoseWarper (Bertasius et al. 2019), we assume the labels are available every 5^{th} frame in the Penn Action and Sub-JHMDB i.e., $K = 5$. We fix the half time-window length of the action space to $T = 2K$ for all datasets, which means the action space of the FSA is $4\times$ the size of the labeled frame interval.

Ablation Study

In this section, we conduct a series of ablation studies on the Sub-JHMDB (Jhuang et al. 2013) dataset to analyze the performance of each component in our proposed REMOTE framework. We adopt the pose estimator in (Xiao, Wu, and Wei 2018) with ResNet50 (He et al. 2016) as our backbone, and gradually add our proposed modules to the backbone.

We first evaluate different frame sampling mechanisms including our proposed dynamic FSA, the fixed distance sampling and random sampling on the Sub-JHMDB dataset. As shown in Table 1, the results of uniformly sampling an unlabeled frame that is n frames away from the labeled frame (MT+Uniform-n) demonstrate that uniformly tracing back for 6 frames (MT+Uniform-6) gives us the best results. This validates our hypothesis that pairing the given labeled frame with either a frame that is too close or too far away is sub-optimal. This is due to the fact that pairing with a frame that is too close (e.g., MT+Uniform-3) brings almost no motion offsets, limiting the amount of new information the MT could capture. While pairing with a frame that is too far away (e.g., MT+Uniform-9) significantly raises the difficulties of conducting motion transformation, leading to a lower PCK values in Table 1. We obtain best performance (78.9) when the unlabeled frame is sampled 6 frames away. However, with our dynamic FSA mechanism, we further outperform it by 1.9 (from 78.9 to 80.8).

In addition to that, we also notice that training on the random pairs of labeled and unlabeled frames (i.e., MT+Random) brings less improvement compared to the best fixed distance sampling schemes (i.e., MT+Uniform-6) but more improvement compared to the sub-optimal ones (i.e., MT+Uniform-3 and MT+Uniform-9). We conjecture the reason is that randomness benefits from the diversity of the information, meanwhile introduces noise, which damages the performance of the MT. More precisely, if frame pairs are selected randomly, then pairs that have *overly large motion offsets* are also fed to the network, making it difficult or even infeasible for MT to perform effective reconstruction, i.e., this brings low-quality or even noisy self-supervision signals for the PE. As shown in Fig. 5 (a), our FSA converges at $20K$ iterations ($\sim$12 hours), while Random selection method converges at $30K$ iterations ($\sim$9 hours). We can see our FSA surpasses Random selection at any iteration, while spending slightly longer training time. We argue this is because the proposed framework has a specific reward set in the FSA, which gives the framework a clear direction towards continuously strengthening the PE, throughout the whole training process.

Then, we investigate the importance of each term in the loss function of our MT (Eq. 11). As shown in Table 1, removing any of the three terms ($\mathcal{L}_{\text{gt}}$, $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{pose}}$) from the loss function of the proposed MT significantly drops the performance of the PE. This means the all three terms are necessary for MT to function well.

Comparison with State-of-the-arts

In this section, we demonstrate the performance of our proposed REMOTE framework by comparing with the SOTA models both quantitatively and qualitatively on the Penn Action and Sub-JHMDB datasets.

Quantitative Comparison. Table 2 reveals the comparison between our REMOTE model trained with frames proposed by FSA and SOTA methods. We use the baseline (Xiao, Wu, and Wei 2018) with ResNet50 (He et al. 2016) as our backbone pose estimator E. Specifically, Song *et al* (Song et al. 2017), LSTM Pose Machine (Luo et al. 2018), DKD (Nie

Dataset	Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Average	#labels/video
Penn Action	Song *et al*(Song et al. 2017)	98.0(-)	97.3(-)	95.1(-)	94.7(-)	97.1(-)	97.1(-)	96.9(-)	96.5(-)	N/A
	LSTM Pose Mach.(Luo et al. 2018)	**98.9**(96.0)	98.6(93.6)	96.6(92.4)	96.6(91.1)	98.2(88.3)	98.2(94.2)	97.5(93.5)	97.7(92.6)	N/A
	DKD(Nie et al. 2019)	98.8(96.6)	98.7(93.7)	96.8(92.9)	97.0(91.2)	98.2(88.8)	98.1(94.3)	97.2(93.7)	97.8(92.9)	N/A
	Baseline(Xiao, Wu, and Wei 2018)	98.1(95.6)	98.2(92.1)	96.3(91.7)	96.4(90.3)	98.4(86.9)	97.5(92.4)	97.1(91.8)	97.4(91.3)	N/A
	PoseWarper(Bertasius et al. 2019)	98.3(95.5)	98.8(92.6)	97.3(92.1)	96.8(90.9)	98.3(87.8)	98.1(93.6)	97.7(92.3)	97.9(91.9)	**13.8**
	KFP (Zhang et al. 2020)(Semi-supervised)	98.5(91.8)	98.4(91.9)	97.1(88.3)	95.2(85.0)	**98.9**(**92.1**)	98.7(91.5)	98.4(90.4)	97.8(90.0)	17.3
	Ours	98.8(**96.8**)	**99.1**(**94.2**)	**97.8**(**94.1**)	**98.1**(**92.3**)	98.7(90.9)	**98.8**(**94.9**)	**98.7**(**94.7**)	**98.6**(**93.8**)	**13.8**
Sub-JHMDB	Song *et al*(Song et al. 2017)	97.1(-)	95.7(-)	87.5(-)	81.6(-)	98.0(-)	92.7(-)	89.8(-)	92.1(-)	N/A
	LSTM Pose Mach. (Luo et al. 2018)	98.2(92.7)	96.5(75.6)	89.6(66.8)	86.0(64.8)	98.7(78.0)	95.6(73.1)	90.9(73.3)	93.6(73.6)	N/A
	DKD (Nie et al. 2019)	98.3(**94.4**)	96.6(78.9)	90.4(69.8)	87.1(67.6)	99.1(81.8)	96.0(79.0)	92.9(78.8)	94.0(77.4)	N/A
	Baseline (Xiao, Wu, and Wei 2018)	97.5(88.5)	97.8(79.5)	91.1(72.3)	86.0(68.9)	**99.6**(84.5)	96.8(78.3)	92.6(78.0)	94.4(77.8)	N/A
	PoseWarper (Bertasius et al. 2019)	97.8(91.3)	97.4(80.9)	91.8(73.6)	90.7(69.2)	97.2(84.8)	97.0(79.7)	94.5(78.6)	95.0(78.8)	**7.2**
	KFP (Zhang et al. 2020)(Semi-supervised)	96.2(84.3)	95.8(82.0)	**94.9**(**77.4**)	**92.9**(**72.4**)	96.0(84.0)	95.4(80.9)	94.4(78.4)	95.2(80.3)	17.5
	Ours	**98.5**(93.8)	**98.1**(**82.9**)	93.7(76.2)	**92.9**(71.3)	97.3(**85.9**)	**97.3**(**81.4**)	**95.2**(**80.3**)	**95.9**(**80.8**)	**7.2**

Table 2: Comparison with the SOTA on the Penn Action and Sub-JHMDB datasets. The scores presented in the table are in the format of $\mathrm{PCK}_{\mathrm{body}}(\mathrm{PCK}_{\mathrm{torso}})$. "-" and N/A indicate the corresponding value is unavailable.

et al. 2019) and Baseline (Xiao, Wu, and Wei 2018) leverage all available pose labels for training. Similar to our model, KFP(Zhang et al. 2020) also introduces a frame proposal-based method. However, the final prediction is interpolated with a pose-dynamics dictionary learned in a self-supervised manner. PoseWarper (Bertasius et al. 2019) was also developed with sparse annotations while the features of the unlabeled frame are only warped to its neighboring labeled frame. We experiment on the similar setting to (Bertasius et al. 2019) for a fair comparison and assuming the pose annotation is available for 13.8 frames per video on average.

As outlined in Table 2, our proposed REMOTE framework outperforms the existing SOTA models trained with the complete set of labels, evaluated under both $\mathrm{PCK}_{\mathrm{body}}$ and $\mathrm{PCK}_{\mathrm{torso}}$ metrics. Using the identical set of labeled frames, our proposed model outperforms PoseWarper (Bertasius et al. 2019) by 0.7 in $\mathrm{PCK}_{\mathrm{body}}$ and 1.3 in $\mathrm{PCK}_{\mathrm{torso}}$ on the Penn Action dataset, owing to the efficient informative frame mining mechanism as well as the additional supervision signals brought by the proposed MT. Compared with the KFP (Zhang et al. 2020), due to inaccurate nature of its interpolation step, our model achieves 3.8 improvement on the Penn Action dataset when evaluated with the more critic $\mathrm{PCK}_{\mathrm{torso}}$.

The results on the Sub-JHMDB dataset are also presented in Table 2, where we observe that our model still outperforms the SOTA models in terms of the both PCK metrics as well as the efficiency of using labeled frames. Note that, compared to KFP, our method uses less than half labeled frames per video (7.2 versus 17.5 labeled frames), but still outperforms KFP by 0.7.

We also investigate the efficiency of the ground truth label usage by examining: 1) the variations of the PCK score when providing more labeled frames; 2) the variations of the PCK score among different methods when using the same amount of labeled frames. We jointly plot the PCK score w.r.t the average number of labeled frames per video in Fig. 5 (b). Compared to the PoseWarper (Bertasius et al. 2019), our model not only outperforms it in all label density settings but also enjoys more boost when given the same amount of extra labeled frames. This indicates the efficiency of our proposed method in exploiting the provided sparse annotations.

Qualitative Results We also present a few qualitative results to visually check if our FSA indeed raises informative frames for training the MT. The selected frames by the proposed FSA are shown in Fig. 6. Given the labeled frames marked with orange bounding boxes, the proposed unlabeled frames are marked with red bounding boxes. We observe a moderate amount of variations in the poses contained in the proposed frames compared to the poses from the labeled frames, which is consistent with our intuition.

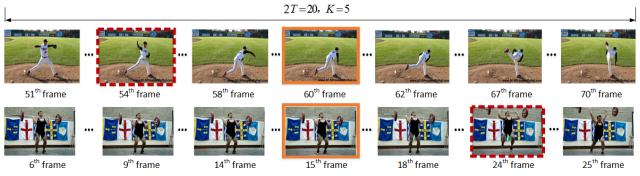

Figure 6: Visualization of the proposed unlabeled frame (red dotted bbox) for the annotated anchor frame (orange bbox). We observe that the selected frames present moderate pose variances, compared to the annotated frame.

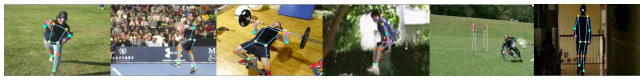

Figure 7: Visualization of the pose estimation results. The left three images are from the Penn Action and the rest are from the Sub-JHMDB.

We further visualize a few pose estimation results in Fig. 7. As depicted in the figure, our model could deliver accurate pose estimation results for different actions in presence of scale/lighting variations.

Conclusion

We address the problem of semi-supervised video human pose estimation in this paper, where only temporally sparse annotations are available. To handle this task, we have proposed the REMOTE framework, a novel model integrating a Motion Transformer (MT) and an RL-based Frame Selection Agent (FSA), that is capable of training the pose estimator based on both labeled frames and temporal dynamics. We conduct extensive experiments that demonstrate the efficacy of the proposed framework.

Acknowledgments

The project is supported by AI Singapore under the grant number AISG-100E-2020-065. The research is also supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andriluka, M.; Roth, S.; and Schiele, B. 2009. Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE conference on computer vision and pattern recognition*, 1014–1021. IEEE.

Belagiannis, V.; and Zisserman, A. 2017. Recurrent human pose estimation. In *IEEE International Conference on Automatic Face & Gesture Recognition*, 468–475.

Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; and Torresani, L. 2019. Learning Temporal Pose Estimation from Sparsely-Labeled Videos. In *Advances in Neural Information Processing Systems*, volume 32, 3027–3038.

Casanova, A.; Pinheiro, P. O.; Rostamzadeh, N.; and Pal, C. J. 2020. Reinforced active learning for image segmentation. *arXiv preprint arXiv:2002.06583*.

Charles, J.; Pfister, T.; Magee, D.; Hogg, D.; and Zisserman, A. 2016. Personalizing human video pose estimation. In *IEEE conference on computer vision and pattern recognition*, 3063–3072.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7103–7112.

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Girdhar, R.; Gkioxari, G.; Torresani, L.; Paluri, M.; and Tran, D. 2018. Detect-and-track: Efficient pose estimation in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 350–359.

Gkioxari, G.; Toshev, A.; and Jaitly, N. 2016. Chained Predictions Using Convolutional Neural Networks. In *European Conference on Computer Vision*, volume 9908 of *Lecture Notes in Computer Science*, 728–743. Springer.

Gui, L.-Y.; Wang, Y.-X.; Liang, X.; and Moura, J. M. 2018. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 786–803.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hernandez, A.; Gall, J.; and Moreno-Noguer, F. 2019. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7134–7143.

Huang, W.; Liang, C.; Yu, Y.; Wang, Z.; Ruan, W.; and Hu, R. 2018. Video-based person re-identification via self paced weighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Insafutdinov, E.; Andriluka, M.; Pishchulin, L.; Tang, S.; Levinkov, E.; Andres, B.; and Schiele, B. 2017. Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6457–6465.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards Understanding Action Recognition. In *IEEE International Conference on Computer Vision*, 3192–3199.

Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Wang, Z.; Wang, X.; Jiang, J.; and Lin, C.-W. 2021. Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining. *IEEE Transactions on Image Processing*, 30: 7404–7418.

Jiang, K.; Wang, Z.; Yi, P.; Lu, T.; Jiang, J.; and Xiong, Z. 2020. Dual-path deep fusion network for face image hallucination. *IEEE Transactions on Neural Networks and Learning Systems*.

Lin, M.; Lin, L.; Liang, X.; Wang, K.; and Cheng, H. 2017. Recurrent 3d pose sequence machines. In *IEEE conference on computer vision and pattern recognition*, 810–819.

Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; and Wang, X. 2021. Deep Dual Consecutive Network for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 525–534.

Luo, Y.; Ren, J. S. J.; Wang, Z.; Sun, W.; Pan, J.; Liu, J.; Pang, J.; and Lin, L. 2018. LSTM Pose Machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5207–5215.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. In *Advances in neural information processing systems*, 406–416.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. A. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR*, abs/1312.5602.

Nie, X.; Li, Y.; Luo, L.; Zhang, N.; and Feng, J. 2019. Dynamic kernel distillation for efficient pose estimation in videos. In *IEEE International Conference on Computer Vision*, 6942–6950.

Pfister, T.; Charles, J.; and Zisserman, A. 2015. Flowing convnets for human pose estimation in videos. In *IEEE International Conference on Computer Vision*, 1913–1921.

Pishchulin, L.; Andriluka, M.; Gehler, P.; and Schiele, B. 2013. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE international conference on Computer Vision*, 3487–3494.

Ruan, W.; Chen, J.; Wu, Y.; Wang, J.; Liang, C.; Hu, R.; and Jiang, J. 2019a. Multi-Correlation Filters With Triangle-Structure Constraints for Object Tracking. *TMM*, 21: 1122–1134.

Ruan, W.; Liu, W.; Bao, Q.; Chen, J.; Cheng, Y.; and Mei, T. 2019b. POINet: Pose-Guided Ovonic Insight Network for Multi-Person Pose Tracking. In *ACM MM*, 284–292.

Ruan, W.; Ye, M.; Wu, Y.; Liu, W.; Liang, C.; Chen, J.; Li, G.; and Lin, C. 2021. TICNet: A Target-Insight Correlation Network for Object Tracking. *TCYB*.

Song, J.; Wang, L.; Van Gool, L.; and Hilliges, O. 2017. Thin-slicing network: A deep structured model for pose estimation in videos. In *IEEE conference on computer vision and pattern recognition*, 4220–4229.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Sutton, R. S. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, 3: 9–44.

Tian, Y.; Zitnick, C. L.; and Narasimhan, S. G. 2012. Exploring the spatial hierarchy of mixture models for human pose estimation. In *European conference on computer vision*, 256–269. Springer.

Toshev, A.; and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1653–1660.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Wang, M.; Tighe, J.; and Modolo, D. 2020. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11088–11096.

Wang, Y.; Chen, Z.; Jiang, H.; Song, S.; Han, Y.; and Huang, G. 2021. Adaptive Focus for Efficient Video Recognition. *arXiv preprint arXiv:2105.03245*.

Wu, W.; He, D.; Tan, X.; Chen, S.; and Wen, S. 2019. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6222–6231.

Wu, Z.; Li, H.; Xiong, C.; Jiang, Y.-G.; and Davis, L. S. 2020. A dynamic frame selection framework for fast video recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple Baselines for Human Pose Estimation and Tracking. In *European Conference on Computer Vision*, volume 11210 of *Lecture Notes in Computer Science*, 472–487. Springer.

Yang, F.; Zhong, Z.; Liu, H.; Wang, Z.; Luo, Z.; Li, S.; Sebe, N.; and Satoh, S. 2021. Learning to Attack Real-World Models for Person Re-identification via Virtual-Guided Meta-Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3128–3135.

Yang, Y.; and Ramanan, D. 2013. Articulated Human Detection with Flexible Mixtures of Parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12): 2878–2890.

Zhang, D.; Guo, G.; Huang, D.; and Han, J. 2018. Poseflow: A deep motion representation for understanding human behaviors in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6762–6770.

Zhang, W.; Zhu, M.; and Derpanis, K. G. 2013. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding. In *IEEE International Conference on Computer Vision*, 2248–2255.

Zhang, Y.; Wang, Y.; Camps, O. I.; and Sznaier, M. 2020. Key Frame Proposal Network for Efficient Pose Estimation in Videos. In *European Conference on Computer Vision*.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zhu, Z.; Huang, T.; Shi, B.; Yu, M.; Wang, B.; and Bai, X. 2019. Progressive Pose Attention Transfer for Person Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2347–2356.