V3D-LSFM:Video-based Human 3D Pose Estimation with Long-term and Short-term Pose Fusion Mechanism

paperId:374, The first author is a student.

No Institute Given

Abstract. Following the success of 2D pose estimation from a single image, a lot of work focus on 3D pose estimation on video by exploiting temporal information. In the scenarios of temporal 3D pose estimation, several recent works have achieved significant advances via Temporal Convolution Network (TCN). However, the current TCN fashion suffers from lacking local coherence caused by excessive dependence on local frames and limited local dynamic range, failing to estimate poses correctly in real scenes, especially with high-speed motions. To tackle this problem, we design a Long-term Bank to select and collect candidate key poses, and further provide a **LSFM**(**L**ong-term and **S**hort-term **F**usion **M**echanism) to integrate long-term pose information into short-term convolution window, such to enhance the temporal coherence of local neighbor frames. Experimental results and ablation studies demonstrate that proposed approaches significantly promote the accuracy and robustness of the benchmark model.

Keywords: 3D Pose Estimation · Video

1 Introduction

Human pose estimation is a classic task in computer vision. In recent years, 2D pose estimation[16, 1, 5, 22, 23, 3] has made significant progress and is gradually matured. The technique are playing critical role in many fields, such as behavior detection in security monitoring, posture correction in medical health, etc. Extending the 2D applications to 3D scenarios, 3D pose estimation are attracting increasing attention in areas 3D motion capture, body reconstruction, animation synthesis, etc.

Although several researchers attempt to learn 3D pose directly from a single image, yet it remains challenges due to the inherent ambiguity of 2D information. In contrast,videos provide additional explicit and implicit temporal constraints for 3D skeleton estimation. To mine more temporal information, a large family of the existing work[19, 13, 4, 20, 10] employs time sequence models of Deep neural network. However, all these methods only use short-term local motions, which are usually discontinuous on high-speed conditions. This challenges are illustrated as in Figure 1. In this case, local poses($L_{t-2}, L_{t-1}, L_t, L_{t+1}, L_{t+2}$) are

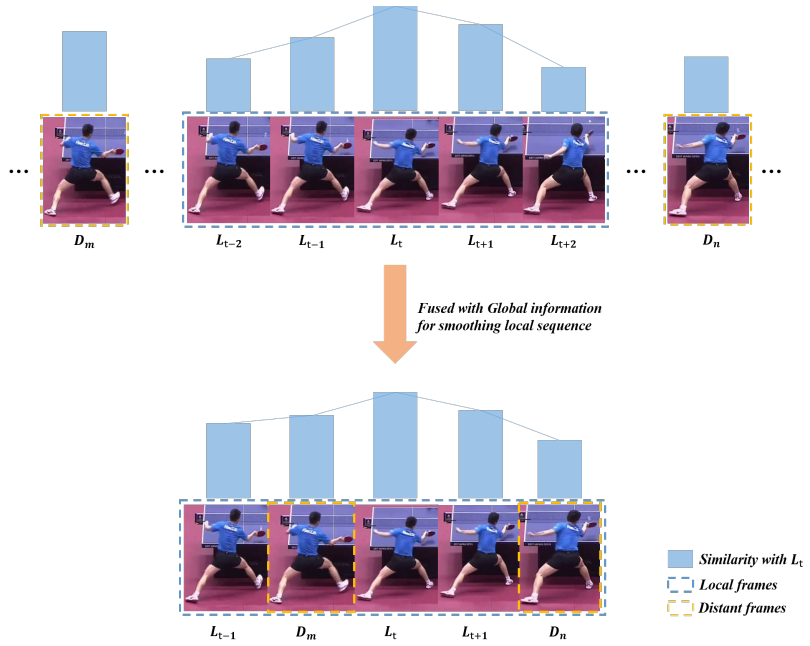

Fig. 1. A **fusion** example in our work. Origin local sequence$(L_{t-2}, L_{t-1}, L_t, L_{t+1}, L_{t+2})$ is fused with distant frames(D_m, D_n) for smoothing itself.

incoherent reflected by the similarity with the center pose L_t. This unsmooth dynamic hinders the performance of convolution in infer phase.

Our method is inspired by the idea of Wu et al.[24], where short-term and long-term features are utilized to capture local and global video dynamics. With this motivation in mind, we designed a long-term and short-term information fusion mechanism. Firstly, we build a Long-term Bank to store keyframes by a greedy strategy and update it dynamically. And secondly, the local poses are reconstructed with keyframes via a similarity-based rearrangement method before local convolution operation.

The contributions of this paper can be summarized:

1. We design a 3D skeleton extraction algorithm that combines long-term and short-term information, which can effectively exploit local and global features to increase local continuity.
2. We propose an effective example-based long-term pose extraction mechanism, and long-term and short-term pose fusion mechanism. Both algorithms are intuitive and high efficiency.
3. We conduct comprehensive experiments to evidence the effectiveness of the proposed methods.

2 Related Work

2.1 2D Pose Estimation from a single Image

In essence, 2D pose estimation is a regression task that learns coordinates of joints position from image pixels. The single-person 2D pose estimation output probabilistic heatmaps of human joints via taking a monocular RGB image as input, and then obtain the final coordinates from the heatmaps. For instance, CPM[23], Stacked Hourglass[16], etc. With the foundation of single-person estimation, the task can be extended to multi-person situations. Multi-person 2D pose estimation can be divided into top-down and bottom-up methods. The top-down refers to first detecting every single person's bounding box, and then performing single-person pose estimation for each person separately, such as AlphaPose[5], CPN[3], HRNet[22]. In contrast, the bottom-up method first estimates the joint points of various parts, and clusters joints of different people together to produce different individuals finally, such as work [2, 15, 6]. 2D human body pose estimation remains far from the practical applications. Thus, quite a lot of work has emerged in the three-dimension domain.

2.2 3D Pose Estimation from a single Image

There are two branches of algorithms for estimating the human body joints position from a single image. One of them is one-stage that directly predicts 3D pose from a monocular image, e.g., work[12, 18]. Another branch requires two steps. Specifically, estimate 2D pose at first, and then lift 2D pose to 3D pose. Powered by existing successful 2D pose detection models, the 3D pose estimation task can be optimized independently, and also the complexity of model is less. Previous work[14, 11] follow this branch. Because of the ambiguity that several 3D poses can be projected to the same 2D pose, several work[8, 17] take leverage of NRSFM(Non-rigid Structure from Motion) to recover 3D from 2D.

2.3 3D Pose Estimation on Video

Compared to a single monocular image, there are more temporal and spatial semantics in videos. The constraints of multi-view, spatial geometric, and temporal continuity can be used to mitigate the impact of self-occlusion and ambiguity. Even though several approaches[21] adopt the End-to-End method to estimate 3D poses from images, the state of the art achievements are made by learning from 2D pose sequences. Different from the work using LSTM[20, 10], Dario et al.[19] employ TCN[9] with stacked dilated convolution layers to predict 3D pose of the center frame by weighting adjacent frames. The accuracy of their model mostly depends on local receptive field size, such as 243 frames in their paper, which is much large for normal videos. When reducing the size, some helpful information will be lost.

To balance both aspects, we design a Long-term Bank to store key poses extracted from the input sequence based on our selection strategy, and further

propose a novel long-term and short-term information fusion algorithm to promote the local coherence.

3 Method Overview

The pipeline of our method is shown as Figure 2. Given a sequence of images, the pipeline first adopt a pretrained 2D pose estimation network to infer the corresponding 2D positions of human joints frame by frame. Based on the estimated per-frame 2D pose, our method sequentially upgrade the temporal 2D poses to 3D poses using Temporal Convocation Network(TCN). During the estimation, our method dynamically maintains a long-term pose bank to store diverse candidate 2D key poses, and adaptively select and integrate coherent long-term poses into the local convolution window (Short-term Bank).

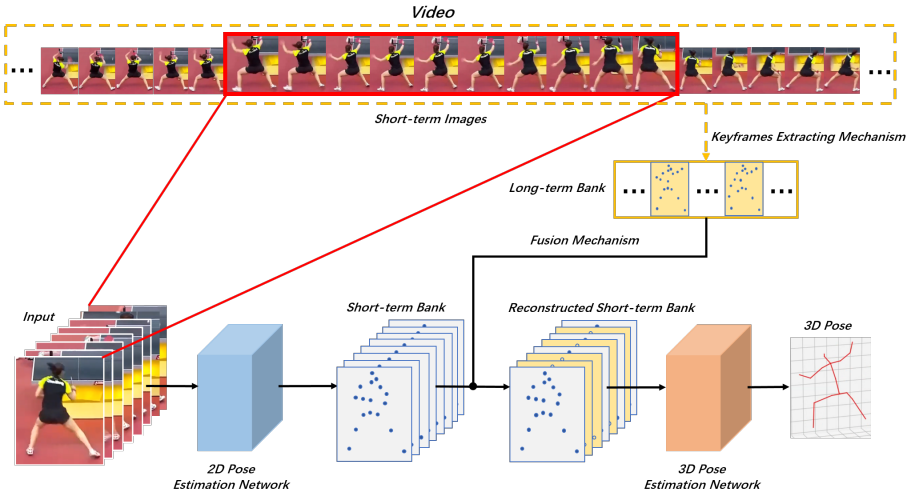

Fig. 2. The pipeline of our method

3.1 3D Pose Estimation Temporal Convolution Network

We employ the fashion of Temporal Convolution Network[19] as the framework of 3D pose estimator, as is illustrated in Figure 3. In this framework, the input is 2D poses $X = \{x_1^0, x_2^0, ..., x_T^0\}$, the model transforms them to a 3D pose of center frame through a series of residual blocks. Each residual block contains a dilation convolution layer with a kernel size of 3 and a linear layer with a kernel size of 1, formulated as

$$x_t^{i+1} = x_t^i + wl^i(w_0^i x_t^i + w_1^i x_{t-d}^i + w_2^i x_{t+d}^i) \quad (1)$$

where d is dilation rate, w^i is convolution weight, and wl^i is linear weight. The TCN model is trained via a semi-supervised manner, where the unsupervised loss ensure the projected estimated 3D poses are consistent with the input 2D poses[19].

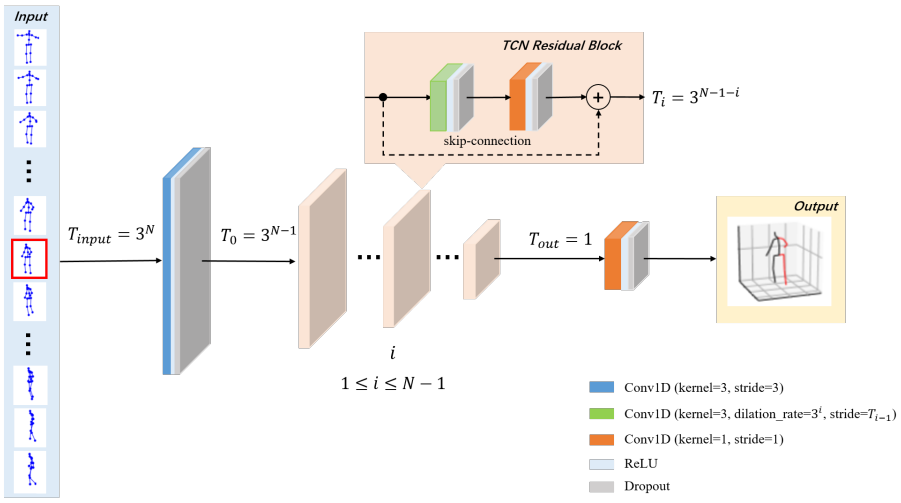

Fig. 3. Structure of Temporal Convolution Network

3.2 Long-term and Short-term Poses Fusion Mechanism

Although Temporal Convolution Network capture smoothly continuous dynamics, it lacks the ability to model discontinuous motions, which commonly happens on fast moving humans. To tackle this shortcoming, we collect a set of the Long-term key poses and integrate them into the short-term poses of the sliding temporal window of TCN, so as to reconstruct the local continuity of the human motion.

Construction of Long-term Pose Bank The Long-term Pose Bank (LTB) is designed to represent motion patterns of the whole sequence by storing key poses. To this end, the poses selected into LTB should be as diverse as possible. Therefore, we devise a similarity-based method to incrementally select diverse poses into the LTB from scratch as the Temporal convolution network proceeds in time sequence. The measurement of similarity between two 2D poses are defined based on their positions of joints:

$$S = 1 - \frac{\sum_{i=0}^{J-1} \|A_{transi} - B_i\|_2}{J \times avg\_bone\_length(B)} \tag{2}$$

where A_{trans} is a translated pose from pose A to pose B by subtracting the offset of their centers, and J is the number of joint.

Based on the measurement of similarity between two poses, we adopt a greedy strategy to progressively insert key poses into LTB. In detail, when Long-term Bank LTB is not full, the candidate frame x is inserted if the similarity S with all stored frames is less than the preset threshold T. While when reaching the max capacity $Size$, we dynamically maintain an accumulated similarity value AS for each keyframe pose and update poses by keeping the sum of AS of the entire bank as little as possible. The process is illustrated as Algorithm 1. In practice, for online video streams, keyframes are collected from the past. As for offline videos, they can also be obtained from the future.

Algorithm 1 Greedy Progressive Insertion Algorithm

Require: candidate frame x, Long-term Bank LTB with capacity $Size$ and similarity threshold T

if $length(LTB) < Size$ **then**
  for each LTB_i **do**
    $S := Similarity(x, LTB_i)$
    if $S < T$ **then**
      insert x into LTB
      update AS for each LTB_i
    end if
  end for
else
  for each LTB_i **do**
    $S := Similarity(x, LTB_i)$
    if $S < T$ **then**
      delete item where AS is max
      insert x into LTB
      update AS for each LTB_i
    end if
  end for
end if

Fusion of Long-term Bank and Short-term Bank The fusion mechanism aims to ensure that the Short-term Bank after fusion is more continuous than before. We design the Alternative Symmetric Pose Fusion Algorithm to dynamically fuse key poses of LTB into STB, targeting at enhancing the motion coherency of the local spatial convolution window. For the center frame of the convolution window of TCN, we firstly calculate the pose similarities between the center frame and 2D key poses that stored in LTB. Then, these similarity value S_c are sorted in the descending order. Then the 2D poses in LTB which are similar with center frame enough (exceed a threshold) are insert into STB. To keep the balance of left and right half side of the STB window, the insertion

is performed in an alternative symmetric manner. An example of the algorithm is shown as Figure 4.

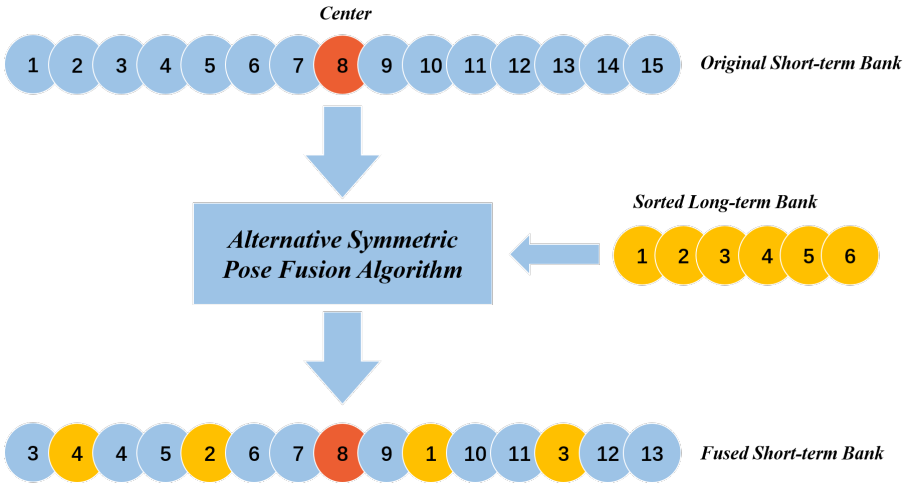

Fig. 4. A fusion example of **Alternative Symmetric Pose Fusion Algorithm**.

3.3 Multiple rate frame sampling for data augmentation

Unlike the smooth motions collected in indoor environments with high-speed cameras, the poses in real scenes usually have multiple dynamic ranges. To encourage the network to learn this pattern, we augment the training data by uniformly sample original continuous sequence with varieties of sampling intervals to synthesize different motion dynamics. In our experiments, the sampling intervals are set to 3, 5, 7.

3.4 Implementation Details

Training. Considering the efficiency of training, we directly train TCN on Human3.6M with ground-truth 2D keypoints. The loss function is the mean per-joint position error between predictions and 3D annotations. Adopting Adam as optimizer with momentum=0.1, we train TCN with a receptive field of 27 frames for 80 epochs. The learning rate and batch size are set to 1e-3 and 1024 respectively.

Inference. The framework has two stages in infer phase. In the first stage, 2D poses are predicted by HRNet, pre-trained on the MS-COCO dataset. In the second stage, before operating convolution in TCN, keyframes are collected into Long-term Bank, and local 2D poses are fused with them. Adopting a 17-joint skeleton, We set capacity size=80 and similarity threshold=0.7 for Long-term Bank.

4 Experiments

4.1 Datasets and Evaluation Protocal

Following previous work[14, 19, 21], we conduct experiments on the Human3.6M dataset[7] as it contains abundant human poses with 2D and 3D annotations and includes 15 common actions. Besides, we use the pseudo-COCO dataset transformed by Human3.6M to evaluate the ability of generalization of our approaches.

In our experiments, we utilize three metrics: MPJPE, P-MPJPE, MPJVE. **MPJPE** is the mean per-joint position error(also used for loss function). **P-MPJPE** is the error after alignment with the ground truth in translation, rotation, and scale. **MPJVE** is the mean per-joint velocity error. Note that all the following experiments are carried out in comparison with the temporal model of the work[19], abbreviated as *VideoPose*. In addition to Human3.6M, we also collect a set of videos with fast moving human, and compare our method with state-of-the-art on these challenging videos.

4.2 Comparison with State-of-the-art

Effectiveness of Long-term Bank Since there exists no large-scale human fast-moving dataset, we build synthetic dataset to evaluate our methods by extracting frames from the Human3.6M origin sequences with a certain interval. With intervals of 3, 5, 7 on testing data, the results in Table 1 and Table 2 show that the model with Long-term Bank outperforms over the *VideoPose* without it. The base model is pre-trained on Human3.6M with a receptive filed of 27. As for the Long-term Bank, the capacity is 80 and the similarity threshold is 0.7.

Effectiveness of Data augmentation We conduct quantified experiments on Human3.6M and pseudo-COCO respectively. Significant results of our method are achieved on both datasets, as is listed in Table 3 and Table 4. Note that the original continuous data is defined as *Sequential*, and the data after frame sampling is denoted as *SampleN*, where N is interval size.

Table 1. Effectiveness of Long-term Bank. MPJPE(mm) results on testing data (TD) with intervals of 3, 5, 7.

TD	LTB	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
3	w/o	51.94	58.51	49.00	58.35	50.45	54.59	55.63	61.09	52.40	58.61	50.87	55.72	60.55	55.61	55.29	55.2
	w	**50.25**	**55.73**	**47.31**	**55.61**	**48.84**	**52.57**	**54.22**	**56.71**	**51.05**	**56.38**	**48.90**	**53.48**	**56.71**	**53.52**	**52.90**	**52.9**
5	w/o	60.05	69.65	55.80	71.25	56.49	64.46	64.22	74.95	60.14	70.81	59.34	65.90	71.98	62.25	61.13	64.6
	w	**56.71**	**63.82**	**53.27**	**65.57**	**54.02**	**60.26**	**61.16**	**66.96**	**57.04**	**66.15**	**55.79**	**61.39**	**65.04**	**58.91**	**57.87**	**60.3**
7	w/o	66.58	79.49	62.09	81.75	61.28	73.46	72.01	85.97	67.41	82.74	65.76	74.92	82.62	76.28	72.96	73.7
	w	**61.80**	**70.76**	**58.10**	**73.50**	**57.71**	**67.14**	**67.21**	**75.53**	**62.62**	**75.27**	**60.78**	**68.35**	**72.22**	**67.29**	**64.71**	**66.9**

Table 2. Effectiveness of Long-term Bank. P-MPJPE(mm) results on testing data (TD) with intervals of 3, 5, 7.

TD	LTB	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
3	w/o	37.64	40.56	37.78	43.05	36.48	38.53	39.09	40.32	37.41	42.13	36.60	39.84	44.11	47.69	45.79	40.5
	w	**36.53**	**39.07**	**36.74**	**41.35**	**35.50**	**37.64**	**38.37**	**38.03**	**36.68**	**41.13**	**35.66**	**38.35**	**42.22**	**46.07**	**43.95**	**39.2**
5	w/o	44.82	47.24	42.79	52.52	40.31	43.62	45.37	48.45	41.33	48.31	41.77	47.35	51.40	53.28	50.39	46.6
	w	**42.29**	**44.00**	**41.28**	**48.93**	**39.02**	**41.71**	**43.54**	**44.13**	**39.93**	**46.17**	**40.16**	**44.23**	**47.71**	**50.93**	**48.28**	**44.2**
7	w/o	50.69	53.40	47.61	60.45	43.40	48.57	51.59	55.12	45.09	54.48	45.73	54.08	58.70	66.58	61.25	53.1
	w	**46.91**	**48.38**	**44.99**	**55.03**	**41.45**	**45.43**	**48.28**	**49.38**	**42.98**	**51.10**	**43.43**	**49.37**	**52.76**	**58.68**	**54.34**	**48.8**

Table 3. Ablation of our data augmentation method on Human3.6M.

Training Data	Testing Data	MPJPE↓	P-MPJPE↓	MPJVE↓
Sequential	Sequential	**39.0**	**28.7**	1.89
	Sample3	55.2	40.5	13.54
	Sample5	64.6	46.6	21.80
	Sample7	73.7	53.1	30.37
Sequential×2	Sequential	39.4	28.7	1.89
	Sample3	53.1	38.9	12.75
	Sample5	60.4	43.9	20.58
	Sample7	67.7	49.4	28.74
Sample3+Sequential(Ours)	Sequential	39.2	28.7	**1.79**
	Sample3			
	Sample5			
	Sample7			
Sample5+Sequential(Ours)	Sequential	39.2	28.7	**1.79**
	Sample3	**39.1**	**28.8**	**7.17**
	Sample5	**39.4**	**29.0**	**10.58**
	Sample7	**40.5**	**29.8**	**14.5**
Sample7+Sequential(Ours)	Sequential	39.2	28.7	**1.79**
	Sample3			
	Sample5			
	Sample7			

Table 4. Ablation of our data augmentation method on pseudo-COCO.

Training Data	Testing Data	MPJPE↓	P-MPJPE↓	MPJVE↓
Sequential	Sequential	**57.0**	**43.7**	3.79
	Sample3	72.3	55.3	16.53
	Sample5	81.1	60.4	24.63
	Sample7	90.2	66.5	32.98
Sequential×2	Sequential	57.4	43.9	**3.77**
	Sample3	70.3	53.9	15.88
	Sample5	77.7	58.1	23.67
	Sample7	85.4	63.3	31.82
Sample3+Sequential(Ours)	Sequential	57.3	43.9	4.38
	Sample3			
	Sample5			
	Sample7			
Sample5+Sequential(Ours)	Sequential	57.3	43.9	4.38
	Sample3	**56.5**	**43.4**	**11.34**
	Sample5	**56.5**	**43.5**	**15.67**
	Sample7	**57.6**	**44.5**	**20.69**
Sample7+Sequential(Ours)	Sequential	57.3	43.9	4.38
	Sample3			
	Sample5			
	Sample7			

4.3 Results of Challenging Videos with Fast Moving Human

We collect 15 videos of four representative sports categories from the Internet with the keywords: "Badminton", "Tennis", "Table tennis", and "Skating". The total length of all the videos is approximately 6 hours. Displayed as Figure 5 and Figure 6, we visualized the comparison results, of which our method performs better than the benchmark model.

5 Conclusion

In this paper, we propose a novel temporal information fusion mechanism that integrates short-term poses with long-term poses to improve the coherence of local pose sequence. Qualitative and quantitative experiments demonstrate that the proposed fusion mechanism can promote the accuracy and robustness of the benchmark model. Furthermore, to improve the generality of the basic TCN model against high dynamic motions, we also provide a data augmentation method that concatenates sequences with different frame rates. The results on the Human3.6M dataset and pseudo-COCO dataset indicate the superiority of this method.

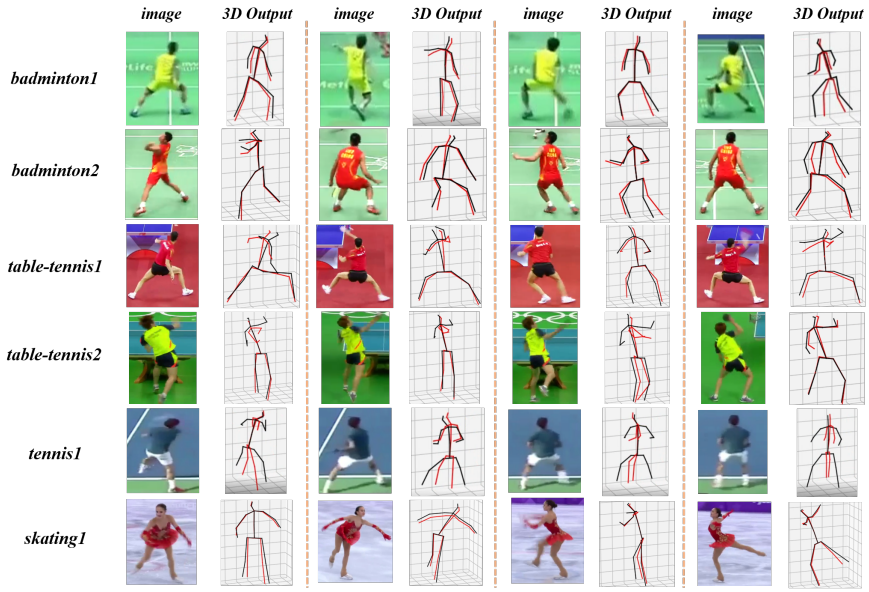

Fig. 5. Qualitative Justification of Long-term Bank. The red skeleton is VideoPose, and the black is the model with *Long-term Bank*.

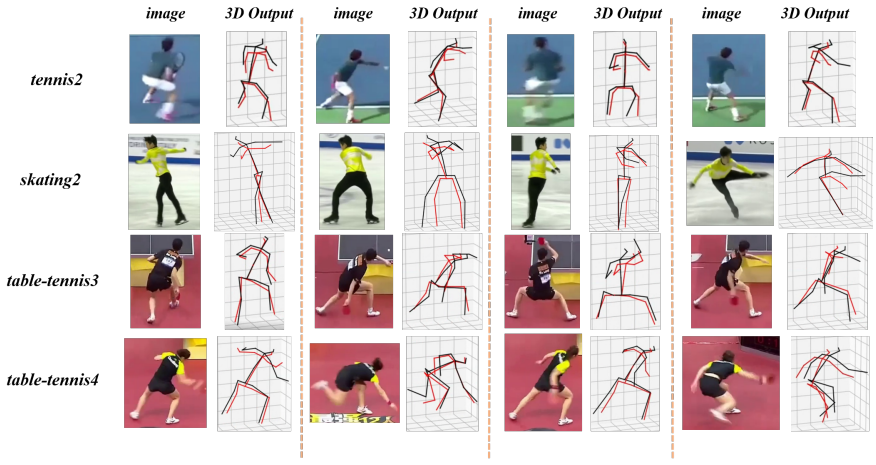

Fig. 6. Qualitative Justification of Multiple Sampling Rate Augmentation. The red skeleton is VideoPose, while the black is trained on augmented data.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. "Densepose: Dense human pose estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7297–7306.

[2] Zhe Cao et al. "Realtime multi-person 2d pose estimation using part affinity fields". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7291–7299.

[3] Yilun Chen et al. "Cascaded pyramid network for multi-person pose estimation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7103–7112.

[4] Rishabh Dabral et al. "Learning 3d human pose from structure and motion". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 668–683.

[5] Hao-Shu Fang et al. "Rmpe: Regional multi-person pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2334–2343.

[6] Gines Hidalgo et al. "Single-Network Whole-Body Pose Estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 6982–6991.

[7] Catalin Ionescu et al. "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments". In: *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013), pp. 1325–1339.

[8] Chen Kong and Simon Lucey. "Deep interpretable non-rigid structure from motion". In: *arXiv preprint arXiv:1902.10840* (2019).

[9] Colin Lea et al. "Temporal convolutional networks for action segmentation and detection". In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 156–165.

[10] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. "Propagating lstm: 3d pose estimation based on joint interdependency". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 119–135.

[11] Chen Li and Gim Hee Lee. "Generating multiple hypotheses for 3d human pose estimation with mixture density network". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 9887–9895.

[12] Sijin Li and Antoni B Chan. "3d human pose estimation from monocular images with deep convolutional neural network". In: *Asian Conference on Computer Vision*. Springer. 2014, pp. 332–347.

[13] Yue Luo et al. "Lstm pose machines". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5207–5215.

[14] Julieta Martinez et al. "A simple yet effective baseline for 3d human pose estimation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2640–2649.

[15] Alejandro Newell, Zhiao Huang, and Jia Deng. "Associative embedding: End-to-end learning for joint detection and grouping". In: *Advances in neural information processing systems*. 2017, pp. 2277–2287.

[16] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation". In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[17] David Novotny et al. "C3DPO: Canonical 3d pose networks for non-rigid structure from motion". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 7688–7697.

[18] Georgios Pavlakos et al. "Coarse-to-fine volumetric prediction for single-image 3D human pose". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7025–7034.

[19] Dario Pavllo et al. "3D human pose estimation in video with temporal convolutions and semi-supervised training". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7753–7762.

[20] Mir Rayat Imtiaz Hossain and James J Little. "Exploiting temporal information for 3d human pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 68–84.

[21] Yu Sun et al. "Human mesh recovery from monocular images via a skeleton-disentangled representation". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 5349–5358.

[22] Jingdong Wang et al. "Deep high-resolution representation learning for visual recognition". In: *arXiv preprint arXiv:1908.07919* (2019).

[23] Shih-En Wei et al. "Convolutional pose machines". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 4724–4732.

[24] Chao-Yuan Wu et al. "Long-term feature banks for detailed video understanding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 284–293.