Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

SPECIAL ISSUE PAPER

SADNet: Generating Immersive Virtual Reality Avatars by Real-time Monocular Pose Estimation

Ling Jiang [1] | Yuan Xiong [1] | Qianqian Wang [1] | Tong Chen [1] | Wei Wu [1] | Zhong Zhou* [1,2]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China

Correspondence
Zhong Zhou, State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and Zhongguancun Laboratory, Beijing, China. Email: zz@buaa.edu.cn

Summary

Generating immersive virtual reality avatars is a challenging task in VR/AR applications, which maps physical human body poses to avatars in virtual scenes for an immersive user experience. However, most existing work is time-consuming and limited by datasets, which does not satisfy immersive and real-time requirements of VR systems. In this paper, we aim to generate 3D real-time virtual reality avatars based on a monocular camera to solve these problems. Specifically, we first design a self-attention distillation network (SADNet) for effective human pose estimation, which is guided by a pre-trained teacher. Secondly, we propose a lightweight pose mapping method for human avatars that utilizes the camera model to map 2D poses to 3D avatar keypoints, generating real-time human avatars with pose consistency. Finally, we integrate our framework into a VR system, displaying generated 3D pose-driven avatars on Helmet-Mounted Display devices for an immersive user experience. We evaluate SADNet on two publicly available datasets. Experimental results show that SADNet achieves a state-of-the-art trade-off between speed and accuracy. In addition, we conducted a user experience study on the performance and immersion of virtual reality avatars. Results show that pose-driven 3D human avatars generated by our method are smooth and attractive.

KEYWORDS:
human pose estimation, computer animation, 3D avatar

1 | INTRODUCTION

Virtual reality avatar generation is favored by the advancement of deep learning and motion capture technologies and has a variety of applications in Virtual Reality. In particular, pose-driven human avatars are urgently demanded to replace time-consuming and expensive hand-crafted [1] virtual characters in telepresence, 3D gaming, augmented and virtual reality communication.

Common methods use a variety of sensors for precise motion capture, including wearable motion capture sensors (digital gloves [2], wearable IMU [3]), multi-view camera systems [4], or expensive 3D scanners [5], and depth cameras [6,7]. These methods are capable of accurately capturing motion signals beyond RGB images with high-cost sensors and can generate human models with high quality. However, these motion capture sensors are generally expensive and time-consuming, while multi-view camera systems require complex environments, which limits their practicability. With the development of deep learning, some studies have started to utilize monocular cameras for pose estimation to drive human avatars and increase their realism. Some of them

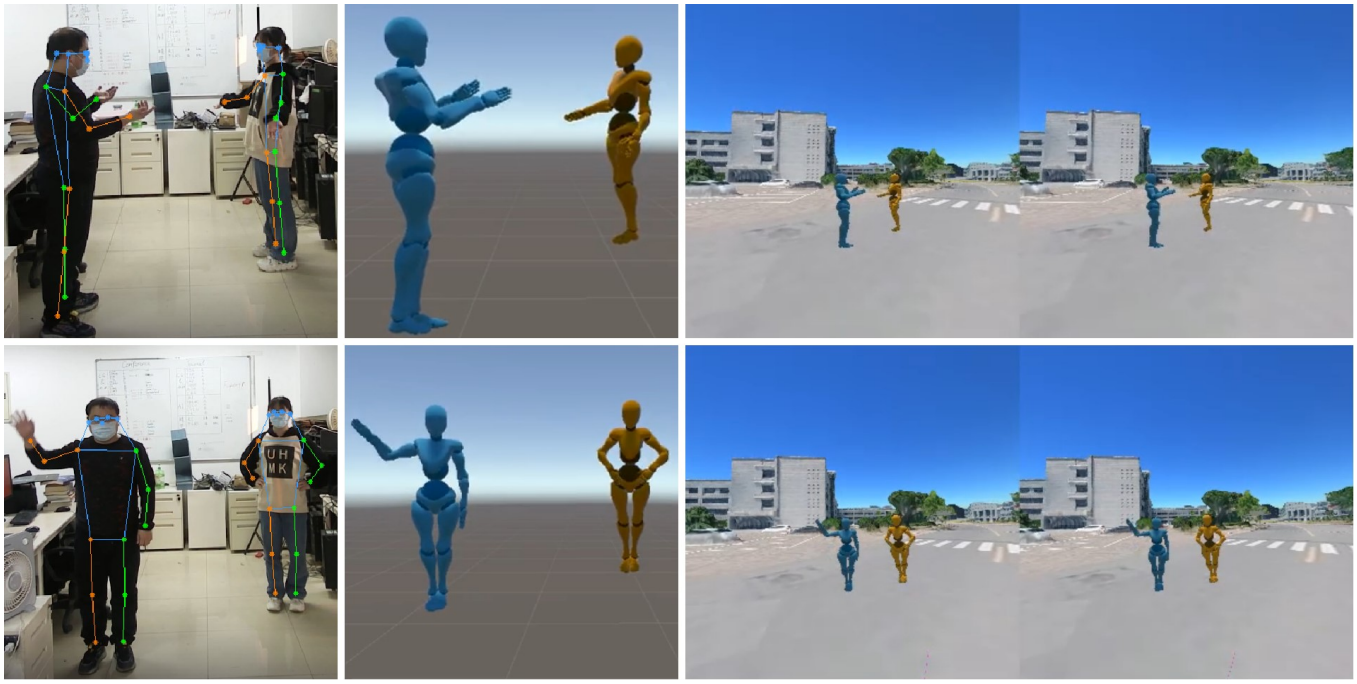

Figure 1 Immersive 3D avatar generation for different scenarios. We first obtain 2D human poses from SADNet (the left column), then generate 3D pose-driven avatars (the middle column). We finally integrate avatars into virtual scenes and display in binocular VR devices for an immersive user experience (the right column).

use deep networks to reconstruct parametric models[8,9] from a single image to obtain human body models with a fixed topology and rough skinning. However, most of them are data-driven. On the one hand, they are trained using fixed datasets that are captured in a homogeneous scene and lack outdoor data. On the other hand, they do not take into account the camera model and environment settings, influencing model robustness. Thus we are motivated to design a lightweight framework that accurately drives human avatars and does not rely on high-cost systems or complex algorithms.

To address the above issues, we generate 3D pose-driven virtual reality avatars based on a monocular camera for presence and immersion requirements. As shown in Figure 1, we first design SADNet for real-time human pose estimation, which employs a teacher-student framework. Secondly, we propose a lightweight pose mapping method for generating human avatars with pose consistency. Finally, we integrate our method into a VR system for an immersive user experience. Comprehensive experimental results show the superiority of our framework at low cost.

Our contributions are summarized as follows:

1. We introduce a unified framework for generating 3D real-time pose-driven human avatars, which presents an immersive experience and connects users with virtual bodies.

2. We propose SADNet for real-time human pose estimation, which uses lightweight litetrans blocks instead of transformer blocks and employs a pre-trained teacher to guide the student through a distillation token.

3. We present a Lightweight pose mapping method for human avatars. It utilizes the camera model to recover 3D human poses and generates dynamic 3D human avatars with pose consistency.

2 | RELATED WORK

Immersive human avatars. Recently, a large amount of work has been devoted to obtaining accurate motion capture using sensors[5,2,4,3,6,7,10]. These methods are capable of accurately capturing motion signals beyond RGB images with expensive sensors, and can generate human models with high quality. However, these motion capture sensors are generally expensive and time-consuming, while multi-view camera systems require complex setup environments, which limits their practicability.

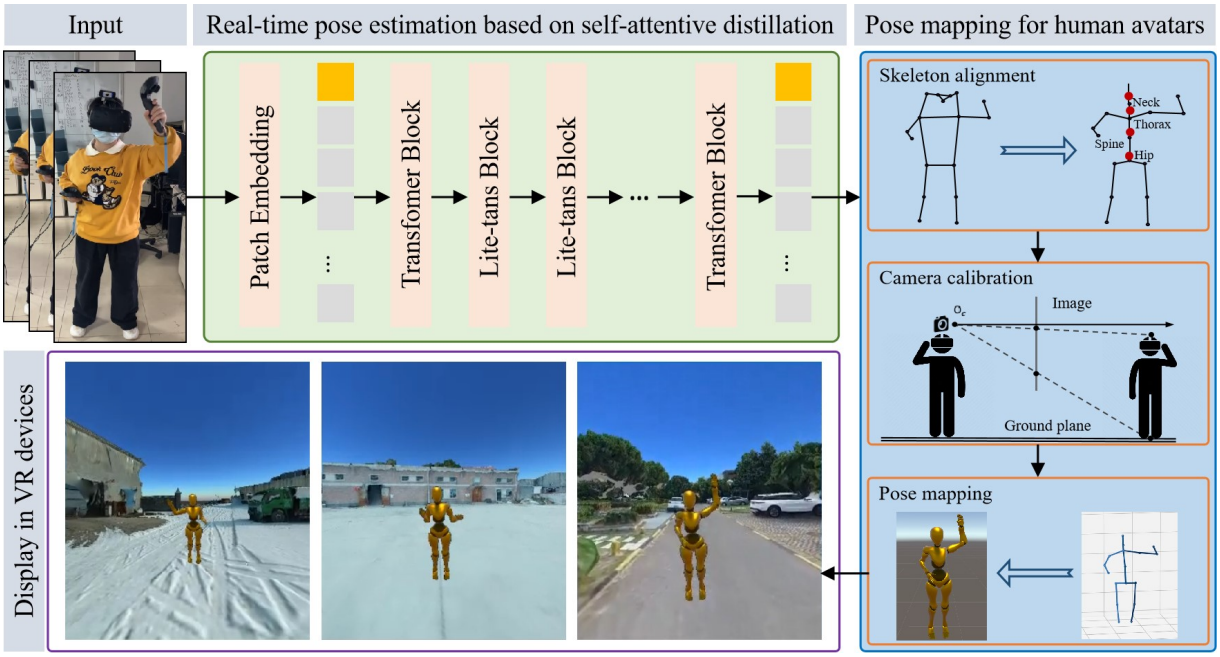

Figure 2 Pipeline of our framework. Firstly, we design SADNet for real-time 2D pose estimation, which uses lightweight litetrans blocks to replace transformer blocks and employs a pre-trained teacher to guide the student. Secondly, we propose an efficient pose mapping method to generate dynamic human avatars with pose consistency. Finally, we integrate avatars into virtual scenes and display in binocular VR devices for an immersive user experience.

With the development of deep learning, some studies design deep networks to regress parametric 3D human models directly from images[8,9,11,12,13,14]. They are divided into two main categories: learning-based and optimization-based methods. Learning-based methods[13,14] refer to regressing the 3D keypoints or meshes of the human body directly from the image. However, these methods are data-driven and receive limitations of the trained dataset. The way of obtaining 3D human keypoint annotations is too complicated, making the data only to be collected in a fixed homogeneous scene and lacking large outdoor data. Optimization-based methods[11,12] usually use reprojection to calculate the error at 2D key points in the image for optimization. These methods are more dependent on the accuracy of 2D human pose estimation and mostly do not consider the physical camera and environment settings, influencing these models' robustness.

Human pose estimation. Human pose estimation is a basic vision task, aiming to detect human keypoints in the given image. Recent studies on 2D pose estimation have achieved excellent performance on public benchmarks[15,16,17,18]. These are divided into two main categories: bottom-up and top-down methods.

Bottom-up methods[15,19,20] first detect all non-identified keypoints and then group them into individual poses. They are more suitable for crowded scenarios and the amount of computation does not vary with the number of people, but need to deal with various human bodies. HrHRNet[15] designs a multi-resolution fusion network to obtain features at different scales for accurately detecting human keypoints. SWAHR[19] proposes adaptive ground-truth human keypoint heatmaps to deal with various human bodies. Pose-AE[20] guides the network to output both group assignment and body joint localization results. However, these methods need to handle a variety of human scales, making them challenging in accuracy and speed of inference.

Top-down methods[16,17,18] first use a human detector to detect the human body and then perform keypoint localization. These methods outperform bottom-up methods in terms of speed and accuracy in non-extreme scenarios (no more than 6 people). TransPose[17] uses CNNs to extract modeling global relationships. TokenPose[18] introduces additional tokens to estimate the location of occluded keypoints and to model relationships between keypoints. ViTPose[16] shows the surprising performance of plain visual transformers for human pose estimation and achieves state-of-the-art performance on public datasets. However, its application in industry still suffers from heavy parameters and high latency. Several methods also work on lightweight models[21,22].

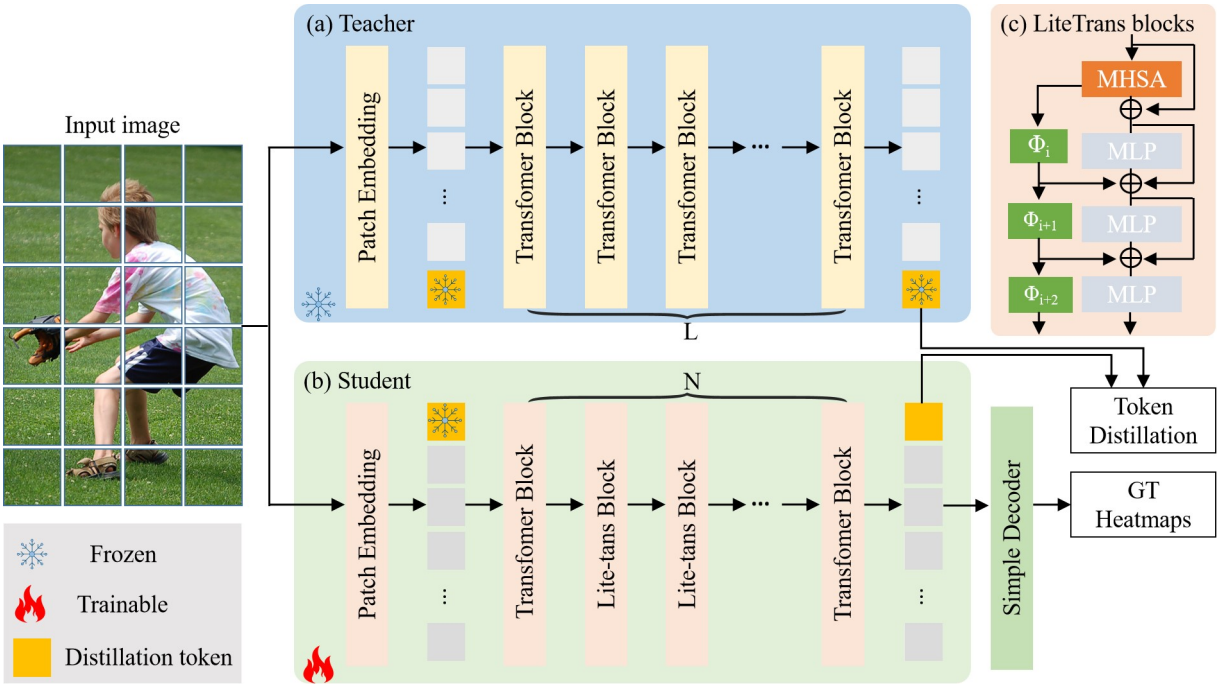

Figure 3 Overview of SADNet. We introduce the teacher-student framework to close the gap between large and small models. The teacher (a) is employed to train the student (b) via distillation loss. (c) LiteTrans block is used to replace self-attention maps with low computing costs.

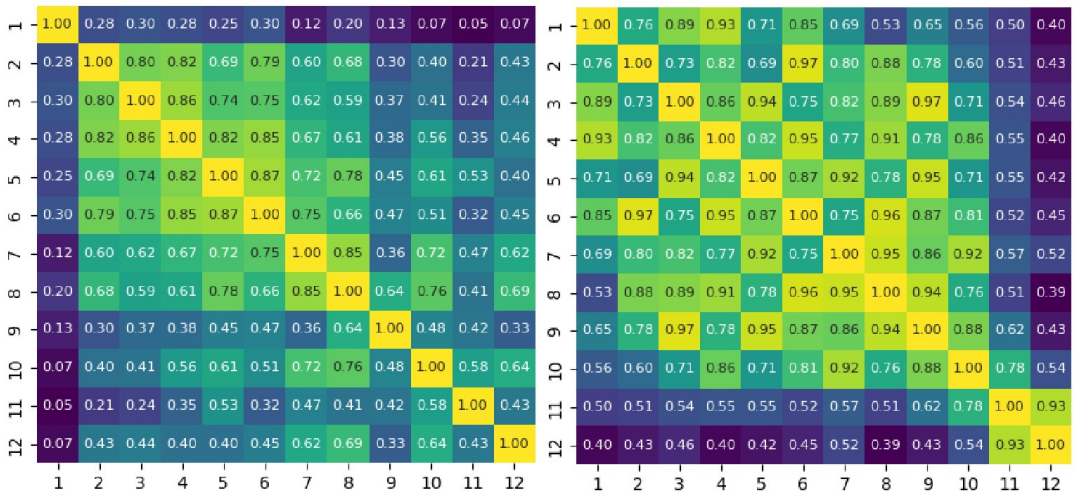

Figure 4 CKA of self-attention maps (a) and MSA (b). ViTPose-S has a high correlation across intermediate 3-8 blocks.

3 | REAL-TIME POSE-DRIVEN HUMAN AVATAR GENERATION

In this paper, we aim to generate 3D pose-driven virtual reality avatars based on a monocular camera for presence and immersion requirements as shown in Figure 2. Firstly, we start with SADNet for real-time 2D pose estimation, which uses lightweight litetrans blocks instead of transformer blocks and employs a pre-trained teacher to guide the student. Secondly, we propose a lightweight pose mapping method for human avatars that generates dynamic human avatars with pose consistency. Finally, we display human avatars in VR devices for an immersive user experience.

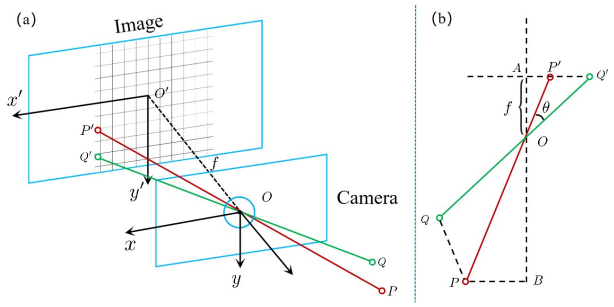

Figure 5 Camera imaging model. Based on 2D pose estimation results, we use the camera imaging model to recover 3D human poses for human avatars.

3.1 | Real-time human pose estimation based on self-attentive distillation

ViTPose[16] shows the surprising performance of plain visual transformers for human pose estimation and achieves state-of-the-art performance. However, the large model parameters with high latency make it difficult to deploy on resource-constrained edge devices. In this paper, we present a real-time pose estimation method based on self-attentive distillation.

Firstly, we introduce Knowledge distillation to compress the model as shown in Figure 3. We employ a pre-trained teacher to guide the student through a distillation token. Specifically, We add a distillation token to the initial embeddings (patches) for feature distillation. Distillation token is used similarly to the class token, except that it aims at reproducing the features estimated by the teacher. This transformer-specific strategy allows our model to learn from the teacher's intermediate features to bridge the gap between large and small models.

The learnable distillation token t is randomly initialized and attached to the visual tokens after the embedding layer of the teacher. Then, the trained teacher is frozen and only the distillation token is updated as Equation 1, where H_{gt} is the groud-truth keypoint heatmaps and X is the input image. $T(t;X)$ denotes output of the teacher, and t^* denotes the optimal token that minimizes the loss.

$$t^* = \arg\min_t(MSE(T(t;X), H_{gt})) \tag{1}$$

During training, the distillation token t^* is frozen and added to the visual tokens in the student network, thus transferring knowledge from the teacher to the student. The losses in the student network are shown in Eq 2, where the first term is the distillation loss of intermediate features, and the second term is the heatmap loss. It is worth noting that the distillation loss occurs only after replaced blocks.

$$\begin{aligned}\mathscr{L} &= MSE(S(t^*;X), F_t)\\ &+MSE(T(t^*;X), H_{gt})\end{aligned} \tag{2}$$

To further compress the model, we study the similarity of intermediate features. It is well established that the self-attention maps (A) and multi-head self-attention blocks (Z) from the class token in the transformer structure exhibit high correlation especially in intermediate layers[23]. Motivated by this observation, we utilize the distillation token to compute the Centered Kernel Alignment (CKA)[24] between the self-attention maps and MSA blocks in ViTPose-S as shown in Figure 4, and find that there is indeed a high degree of similarity in intermediate 3-8 layers, verifying our hypothesis. Therefore we introduce *skipatt*[23] to create lightweight LiteTrans blocks as Figure 3(c).

3.2 | Pose mapping for human avatars

Based on the system service setup, we reasonably assume that the initial state of the target person is to stand naturally facing the camera and the starting motion direction is always towards the camera. Here we propose a lightweight pose mapping method to generate human avatars with pose consistency which are displayed in Helmet-Mounted Display (HMD).

Skeleton alignment. Firstly, we quickly detect human keypoints in the given image by SADNet. Since our model is trained on a fixed dataset, to align with the topology of the human skeleton activity, we crop and correspondingly interpolate the obtained 2D keypoints. We design a human skeleton as shown in Figure 2, where we drop keypoints like eyes, ears and add keypoints

like neck, chest, spine, hip. Besides, our model is designed based on images, so it is obvious that it produces non-smoothness in videos. Thus we use Kalman[25] for de-jittering.

Camera calibration. We bind a camera to HMD, the intrinsic parameters of the camera (M) are pre-obtained, and through the HMD, we can easily obtain the extrinsic parameters of the camera (R rotation matrix, t translation matrix). The distance (Z_m) between the initial plane of the target person and the camera plane can also be obtained by the RoomScale localization system. In addition, we need to pre-enter body scales (shank length, etc.) of the target person.

$$\begin{bmatrix} X \\ Y \\ Z_{const} \end{bmatrix} = R^{-1}M^{-1}Z_m \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} - R^{-1}t \tag{3}$$

3D human pose recovery. We start to recover 3D human keypoints from 2D poses through the camera imaging model as shown in Figure 5(a). At a fixed depth (Z_{const}), we first estimate 3D human keypoints on the initial body plane. We take the right shoulder (P) as an example, where P is a keypoint in the given image detected by SADNet. We can recover the 3D spatial positions of human keypoints according to the camera imaging model in Equation 3. Then we estimate 3D human keypoints outside the initial body plane according to the assumption that the target person moves toward the camera. We take the right elbow (Q) as an example, where Q is a keypoint in the given image detected by SADNet. As shown in Figure 5(b), we can easily calculate the angle θ by the two points P, Q (obtained by SADNet) in the image. In ΔOPQ, we can calculate $\overrightarrow{OQ}$ by Equation 4, then further get the 3D spatial position of Q. Where $|PQ|$ is the length of the right elbow and right shoulder of the target person recorded in advance. It is worth noting that we add motion estimation information such as velocity and acceleration to to deal with the uncertainty of depth in the subsequent motion. We correct for some of the distorted poses by the coherence of the motion between sequential frames. In addition, the RoomScale localization system is employed for pose calibration and global 3D localization, and Kalman[25] is still employed after obtaining 3D human keypoints.

$$\begin{aligned} |OQ|^2 = |OP|^2 + |PQ|^2 - 2|OP|^2\sin^2\theta \\ \pm 2|OP|\sqrt{(|PQ|^2 - |OP|^2\sin^2\theta)\cos^2\theta} \end{aligned} \tag{4}$$

Pose mapping. After obtaining 3D keypoint positions of the target person, we need to calculate the rotation angles between joints to drive the human avatar to move consistently with the target person. Here we mainly calculate the rotation matrix according to the Euler angles. We first build a tree structure with the hip as the root node and then calculate the rotation angle of the current joint node concerning its parent node according to the 3D human body pose to build a human avatar with pose consistency. We can reduce rendering time by just entering the structural data of the model's motion instead of rendering screens. Inter-frame interpolation (spherical linear interpolation) is then performed in the time domain, which solves the problem of lower motion capture speeds not being able to match higher rendering frame rates.

4 | EXPERIMENTS

4.1 | Performance Assessment

Experiment setup. Our framework is integrated into a VR system with C++. We use HMD devices (VIVE Cosmos and VIVE) for experiments. The experimental computer with a 1080ti GPU and Windows 10 system connects HMDs and is configured with SADNet (onnx) and the pose mapping module.

Data flow. We access the online camera and pass images into SADNet model for 2D pose estimation (2D SADNet), use the pose mapping module to activate the human avatar for pose consistency (PoseMap), and finally, display in HMDs(Render).

Performance. We count the average latency required for each module in the computer as shown in Figure 6. It can be seen that the pose estimation takes the longest time in the end-to-end system latency, which can be weighed against time and accuracy when selecting a model. In addition, all our models can reach 60+FPS (frames per second). In particular, we calculate the latency of Ours-S (i7-8700 CPU) model and get 14.69ms, which also can reach 60+FPS. Notably, we perform interframe interpolation (spherical linear interpolation) in the time domain, which solves the problem of lower pose estimation speeds not being able to match higher rendering frame rates.

Visualization in Unity. We visualize the pose-driven human avatars generated by our method in Unity as shown in Figure 7. The first column shows the input image sequences with 2D pose. The other columns show pose sequences of the human avatar

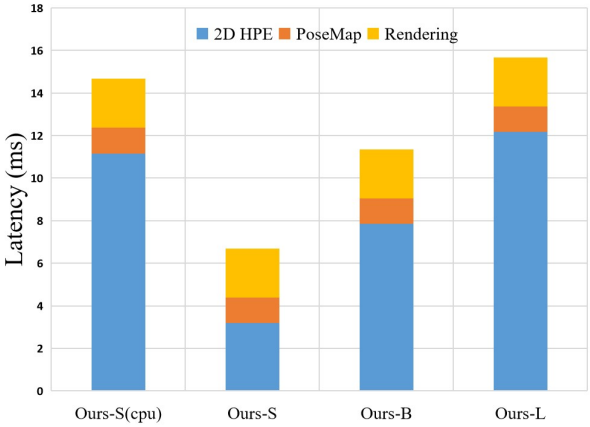

Figure 6 End-to-end system latency.

for different behaviors (sitting, walking, and playing). It can be seen that our model can effectively generate pose-driven avatars under daily behaviors.

4.2 | Comparisons with State-of-the-art Methods

Experimental details of SADNet. Following [16], SADNet is conducted in Ubuntu with a 3090 GPU with two publicly available datasets (MS COCO2017 [26] and MPII [27]). For MS COCO dataset, our models are trained on train2017 (118k images), evaluated on val2017 (5k images), and tested on test-dev in terms of average precision (AP) and average recall (AR). For MPII dataset, our models also are trained on the trainset, and evaluated on the validset using mean average precision (mAP). Unless otherwise stated, the settings of our body detectors follow [16]. We present three models with detailed structures as shown in Table 1.

Table 1 Detail structures about SADNet.

Method	Params	Transformer block	LiteTrans block	Decoder
Ours-S	15.3M	4	4	Simple
Ours-M	60.0M	4	4	Simple
Ours-L	182.1M	4	10	Simple

Results on COCO val set. The performance of our model on COCO val2017 is shown in Table 2. Under similar or lower computing costs, our method is superior to other products. For example, SADNet can achieve 75.9AP at lower FLOPs, which is 0.8AP higher than TokenPose [18]. With similar accuracy, our method is faster than other methods. For example, SADNet achieves the best performance (78.4AP) with faster speed as shown in Figure 8.

Results on COCO test-dev. The performance of SADNet on COCO test-dev is shown in Table 3. Compared to bottom-up methods, SADNet outperforms HrHRNet [21] in all non-extreme scenarios (77.2% vs. 66.4%). Bottom-up methods generally use large-resolution images as input, which increases computing costs. Compared with top-down methods, SADNet achieves a state-of-the-art trade-off between speed and accuracy.

Results on MPII val set. To further evaluate our models, we compared SADNet with SOTA methods on MPII val set with ground truth bounding boxes. Following default settings of MPII, we use PCKh as the evaluation metric of performance. Note that ViTPose* [16] refers to the model trained using MPII train set and extra data, but our models are trained only with MPII train set. The experimental results show that our method reduces parameters and computing costs by nearly half without decreasing the accuracy (93.9 mAP).

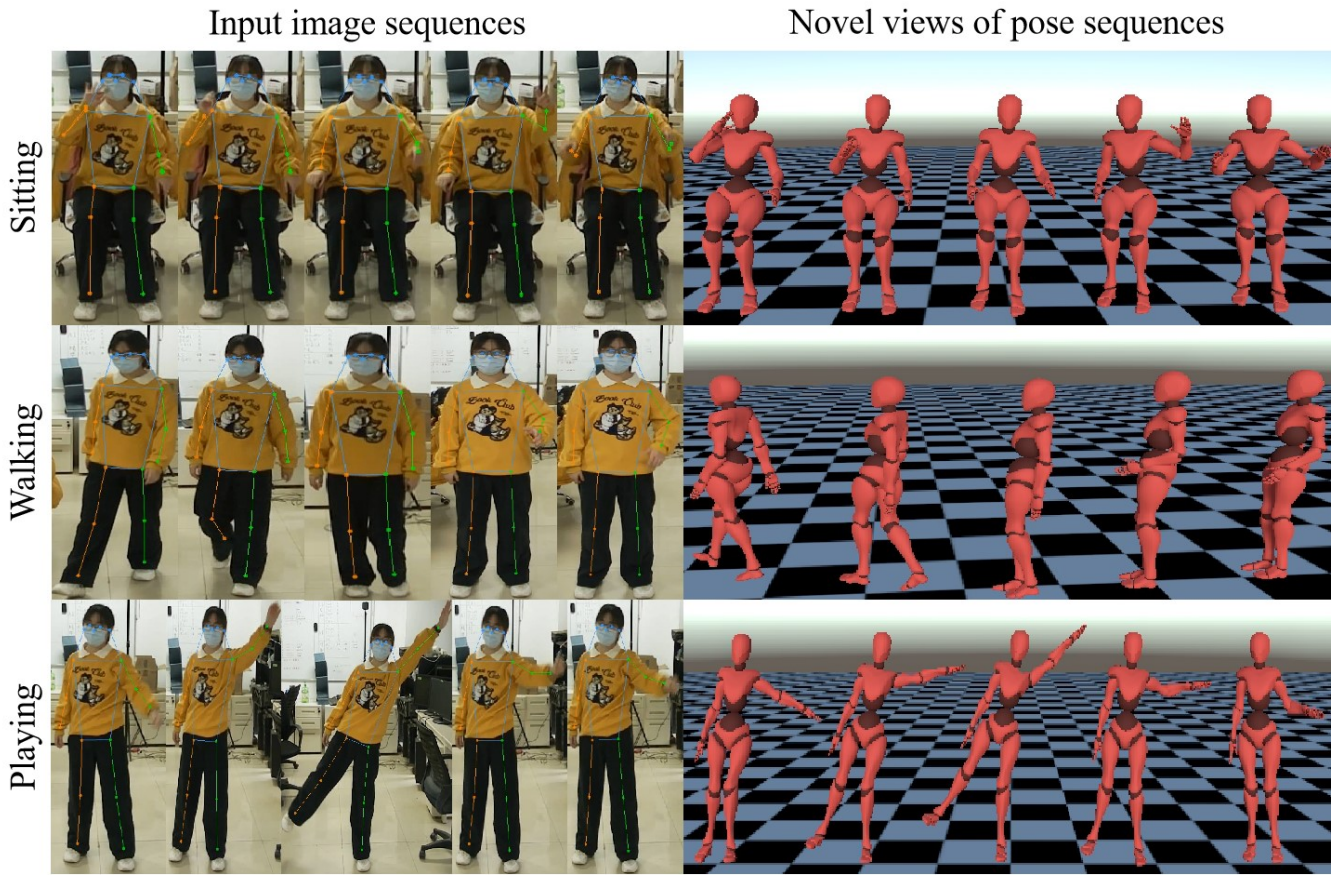

Figure 7 Visualization of different behaviors (sitting, walking, playing) in Unity. We get human pose sequences by SADNet (left) and then generate the human avatar with pose consistency (right). Our models can effectively generate 3D avatars under daily behaviors.

Ablation study. We propose SADNet to perform real-time 2D pose estimation including LiteTrans block and distillation network. To verify the impact of each module, we design ablation experiments using SADNet-B to distill ViTPose-L [16] on COCO val set as shown in Table 5. (a) SADNet-B is a lightweight model of ViTPose-B. (b) After we replaced the Transformer block with the LiteTrans block, the accuracy of SADNet-B decreased by 3.3AP although parameters and computing costs decreased. (c) To close the gap between large and small models, we employ the ViTPose-L model to teach our model for human pose estimation, which brings an improvement of 0.5AP. (d) To improve the accuracy of (b), we propose to use a teacher-student network to address the accuracy degradation caused by the LiteTrans block. Overall, experimental results show that SADNet achieves a state-of-the-art trade-off between speed and accuracy.

4.3 | Application

We integrate the proposed framework into a VR system and display pose-driven human avatars in VR devices for an immersive experience. We get higher smoothness and save system capacity by transferring motion data instead of rendered screens. With motion interpolation, the rendering frame rate on VR devices can reach 60+FPS.

We render pose-driven human avatars into virtual scenes to provide interactive applications. We design two patterns to use our systems: single-user mode and multi-user mode. Under the single-user mode, we capture the human body in the field of view through a monocular camera fixed on the HMD and generate the pose-driven human avatar to be displayed in the user's virtual scene. For multi-user mode as shown in Figure 10, we need multiple users to be equipped with HMDs. They can see human bodies in the field of view in the virtual scene.

Table 2 Comparisons with SOTA methods on COCO val set.

Methods	Input size	Params	GFLOPs	AP	AR
Bottom-up methods					
HrHRNet-W32[21]	512×512	28.6M	47.9	67.1	71.8
HrHRNet-W48[21]	640×640	63.8M	154.3	69.8	76.4
Top-down methods					
HRNet-W32[28]	384×288	28.5M	16.0	75.8	81.0
HRNet-W48[28]	384×288	63.6M	32.9	76.3	81.2
HRNet-W32[28]	256×192	28.5M	7.1	74.4	78.9
HRNet-W48[28]	256×192	63.6M	14.6	75.1	80.4
HRFormer-T[29]	256×192	2.5M	1.3	70.9	76.6
HRFormer-S[29]	256×192	7.80M	2.8	74.0	79.4
HRFormer-S[29]	384×288	7.80M	6.2	74.5	79.8
HRFormer-B[29]	256×192	43.2M	12.2	75.6	80.8
HRFormer-B[29]	384×288	43.2M	26.8	76.2	81.2
TransPose-H-A6[17]	256×192	17.5M	21.8	75.8	80.8
TokenPose-B[18]	256×192	13.5M	5.7	74.0	79.1
TokenPose-L/D24[18]	256×192	27.5M	11.0	75.1	80.2
TokenPose-L/D24[18]	384×288	29.8M	22.1	75.9	80.8
RTMPose-M[30]	256×192	24.7M	1.9	73.6	-
RTMPose-L[30]	256×192	52.3M	4.2	74.8	-
ViTPose-S[16]	256×192	22.0M	5.3	73.8	79.2
ViTPose-B[16]	256×192	86.0M	17.1	75.8	81.1
ViTPose-L[16]	256×192	307.0M	59.8	78.3	83.5
Ours-S	256×192	15.3M	2.9	73.6	79.0
Ours-B	256×192	60.0M	11.5	75.9	81.2
Ours-L	256×192	182.1M	35.1	78.4	83.6

4.4 | User study

We conducted a user experience study with 50 participants to evaluate our method. We compared it with ThreeDPoseTracker[31], VTuber[32] and MW[33]. The user study is designed in terms of smoothness, accuracy, enjoyment, learnability, preference, and latency. We quantitatively built evaluations and experiences into 10 levels, from 1 (“no satisfaction") to 10 (“high satisfaction"). Before and after each experience, participants were asked to fill out the Simulator Sickness Questionnaire[34].

We recruited 50 volunteers from universities and office buildings, including 25 males and 25 females. All of them had no background in professional knowledge as well as experience in animation work. Each participant took 15 minutes to complete this survey. We invited participants for an interactive experience with common activities, such as playing, and dancing.

Based on the total scores for all categories in Figure 9, our method achieved the highest mean score with low standard deviation ($\rho_{ThreeDPoseTracker} = 6.28 \pm 1.25$, $\rho_{KTuber} = 5.70 \pm 2.17$, $\rho_{MW} = 6.60 \pm 1.64$, $\rho_{Ours} = 7.74 \pm 1.09$). Note that no participant illnesses were reported for all tests. Due to the small sample set, it was not appropriate to use the chi-square test directly. The average scoring data shown in Figure 9 was converted into a matrix with 6 columns (scoring criteria) and 4 rows (methods). We calculated the test statistic to get $Q = 22.61$. From the Friedman test table, we can find that it is significant when the upper critical value $F_{0.05}[4, 6] = 7.6$, and the p-value($p < 0.05$). We then utilized the Niemeny test for postoperative comparisons. We calculated the critical distance for average ranking ($CD_{0.05} = 1.915$) with $q_{\alpha=0.05}[4] = 2.569$. It shows that our framework gets a significant difference from 3DPoseTracker and VTuber, and is slightly better than MW. The user study results show that 3D human avatars generated by our method are more attractive.

Table 3 Comparisons with SOTA methods on COCO test-dev.

Methods	Input size	Params	GFLOPs	*AP*	*AR*
Bottom-up methods					
Pose-AE[20]	512×512	277.8M	206.9	58.0	66.1
HrHRNet-W32[21]	512×512	28.6M	47.9	66.4	72.1
Top-down methods					
HRNet-W32[28]	384×288	28.5M	16.0	74.9	80.1
HRNet-W48[28]	384×288	63.6M	32.9	75.5	80.5
HRFormer-S[29]	384×288	7.80M	6.2	74.5	79.8
HRFormer-B[29]	384×288	43.2M	26.8	76.2	81.2
TransPose-H-A6[17]	256×192	17.5M	21.8	75.0	-
TokenPose-B[18]	256×192	13.5M	5.7	74.0	79.1
TokenPose-L/D24[18]	384×288	29.8M	22.1	75.9	80.8
ViTPose-B[16]	256×192	86.0M	17.1	75.1	80.3
ViTPose-L[16]	256×192	307.0M	59.8	77.3	82.4
Ours-B	256×192	60.0M	11.5	75.2	80.4
Ours-L	256×192	182.1M	35.1	77.2	82.1

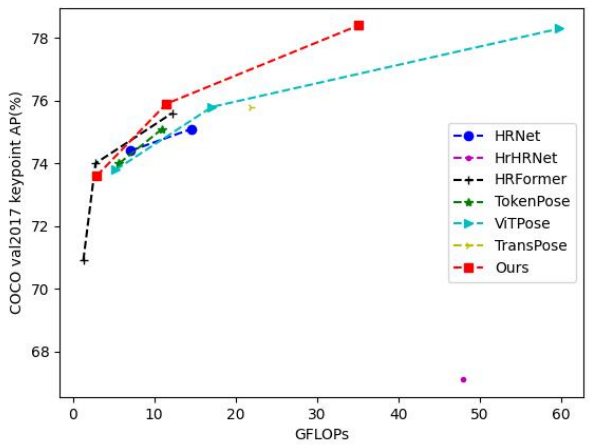

Figure 8 Speed-accuracy performance comparisons on COCO val set. SADNet achieves a state-of-the-art trade-off between speed and accuracy.

5 | CONCLUSION

In this paper, we generate 3D pose-driven virtual reality avatars based on a monocular camera for presence and immersion requirements. As shown in Figure 1, we first design SADNet for real-time human pose estimation, which employs a teacher-student framework. Secondly, we propose a lightweight pose mapping method for generating human avatars with pose consistency. Finally, we integrate our method into a VR system for an immersive user experience. Comprehensive experimental results show the superiority of our framework at low cost.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China under Grant No. 62272018.

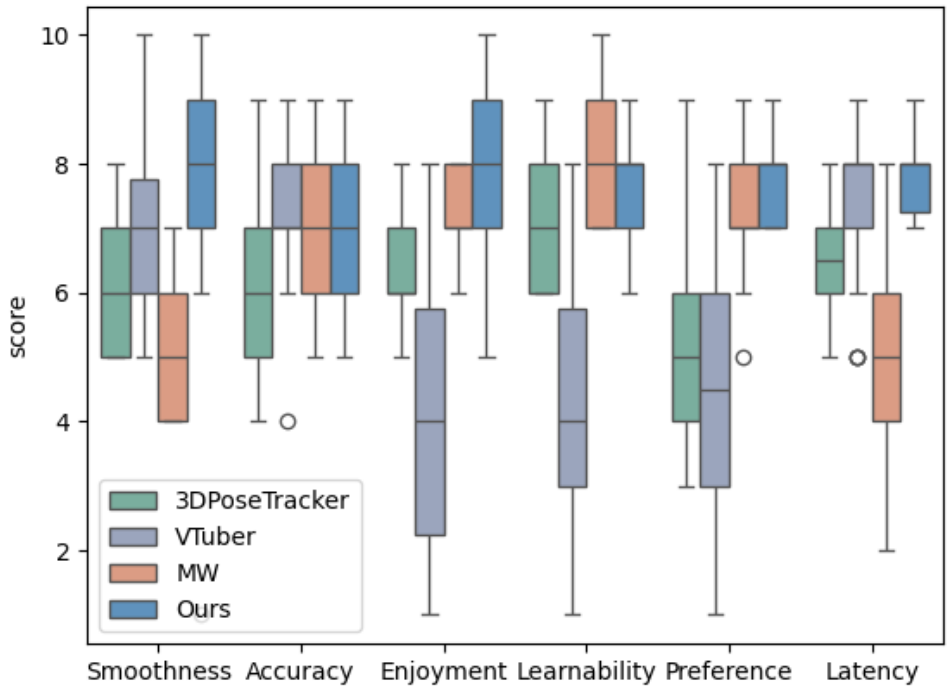

Figure 9 Scoring boxplots of user experience study results.

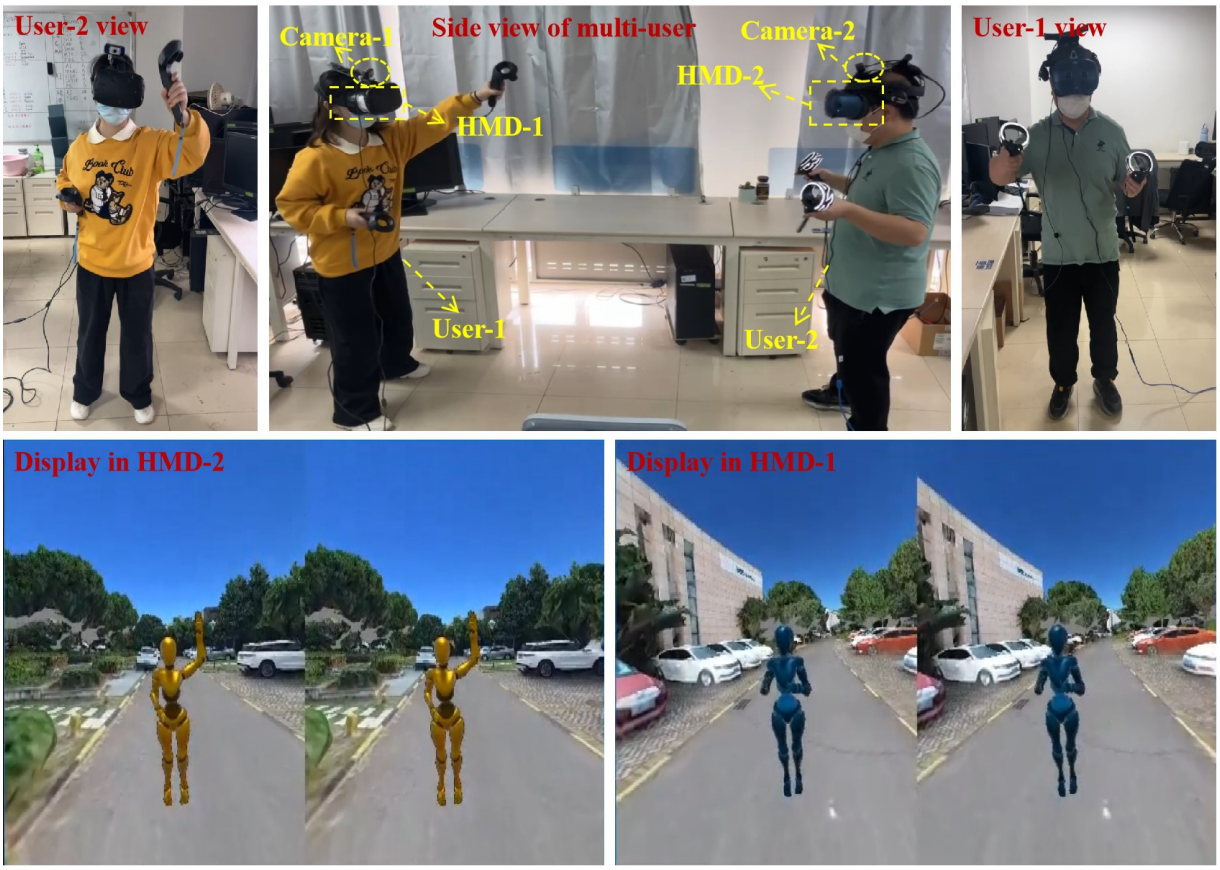

Figure 10 Display in VR devices. In multi-user mode, we use the camera to capture the human body and display the human avatar with pose consistency in the virtual scene.

Table 4 Comparisons with SOTA methods on MPII val set.

Methods	Input size	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
HRNet-W32 [28]	256×256	96.9	85.9	90.5	85.9	89.1	86.1	82.5	90.0
HRNet-W48 [28]	256×256	97.1	95.8	90.7	85.6	89.0	86.8	82.1	90.1
HRFormer-S [29]	256×256	97.1	95.8	90.5	85.9	88.7	85.7	82.1	89.9
HRFormer-B [29]	256×256	96.8	96.7	90.4	85.9	89.0	87.3	84.1	90.4
TransPose-H-A6 [17]	256×192	-	-	-	-	-	-	-	92.3
ViTPose-S* [16]	256×192	97.4	97.2	92.9	89.0	82.3	90.4	86.8	92.7
ViTPose-L* [16]	256×192	98.0	97.6	94.3	90.9	92.9	92.6	89.5	94.0
Ours-S	256×192	97.1	97.2	92.7	89.1	82.3	90.3	86.7	92.3
Ours-B	256×192	97.2	97.3	93.1	89.7	91.5	90.4	87.5	92.5
Ours-L	256×192	97.9	97.6	94.3	90.7	92.9	92.6	89.2	93.9

Table 5 Ablation study on each module.

	LiteTrans	Distillation	*AP*	*AR*
(a)			75.8	81.1
(b)	✓		72.5	78.2
(c)		✓	76.3	81.4
(d)	✓	✓	75.9	81.2

DATA AVAILABILITY

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

References

1. Tang MT, Zhu VL, Popescu V. Alterecho: Loose avatar-streamer coupling for expressive vtubing. In: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE. ; 2021: 128–137.

2. Smith B, Wu C, Wen H, et al. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (TOG)* 2020; 39(6): 1–14.

3. Sharma A, Rombokas E. Improving imu-based prediction of lower limb kinematics in natural environments using egocentric optical flow. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 2022; 30: 699–708.

4. Zhang Y, Li Z, An L, Li M, Yu T, Liu Y. Lightweight multi-person total motion capture using sparse multi-view cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ; 2021: 5560–5569.

5. Cudeiro D, Bolkart T, Laidlaw C, Ranjan A, Black MJ. Capture, learning, and synthesis of 3D speaking styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. ; 2019: 10101–10111.

6. Shao R, Zhang H, Zhang H, et al. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. ; 2022: 15872–15882.

7. Yu T, Zheng Z, Guo K, Liu P, Dai Q, Liu Y. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. ; 2021: 5746–5756.

8. Li Z, Yu T, Zheng Z, Guo K, Liu Y. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. ; 2021: 14162–14172.

9. Kocabas M, Athanasiou N, Black MJ. Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. ; 2020: 5253–5263.

10. Song W, Wang X, Gao Y, Hao A, Hou X. Real-time expressive avatar animation generation based on monocular videos. In: 2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). IEEE. ; 2022: 429–434.

11. Li W, Liu H, Tang H, Wang P, Van Gool L. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. ; 2022: 13147–13156.

12. Shan W, Liu Z, Zhang X, Wang S, Ma S, Gao W. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: European Conference on Computer Vision. Springer. ; 2022: 461–478.

13. Sun X, Xiao B, Wei F, Liang S, Wei Y. Integral human pose regression. In: Proceedings of the European conference on computer vision (ECCV). ; 2018: 529–545.

14. Zhou K, Han X, Jiang N, Jia K, Lu J. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. ; 2019: 2344–2353.

15. Cheng B, Xiao B, Wang J, Shi H, Huang TS, Zhang L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. ; 2020: 5386–5395.

16. Xu Y, Zhang J, Zhang Q, Tao D. ViTPose++: Vision Transformer for Generic Body Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2023.

17. Yang S, Quan Z, Nie M, Yang W. Transpose: Keypoint localization via transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ; 2021: 11802–11812.

18. Li Y, Zhang S, Wang Z, et al. Tokenpose: Learning keypoint tokens for human pose estimation. In: Proceedings of the IEEE/CVF International conference on computer vision. ; 2021: 11313–11322.

19. Luo Z, Wang Z, Huang Y, Wang L, Tan T, Zhou E. Rethinking the heatmap regression for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. ; 2021: 13264–13273.

20. Newell A, Huang Z, Deng J. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems* 2017; 30.

21. Yang Z, Zeng A, Yuan C, Li Y. Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ; 2023: 4210–4220.

22. Wang Y, Li M, Cai H, Chen WM, Han S. Lite pose: Efficient architecture design for 2d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. ; 2022: 13126–13136.

23. Venkataramanan S, Ghodrati A, Asano YM, Porikli F, Habibian A. Skip-Attention: Improving Vision Transformers by Paying Less Attention. *arXiv preprint arXiv:2301.02240* 2023.

24. Kornblith S, Norouzi M, Lee H, Hinton G. Similarity of neural network representations revisited. In: International conference on machine learning. PMLR. ; 2019: 3519–3529.

25. Simon D. Optimal State Estimation, Kalman and Nonlinear Approaches, A John Wiley & Sons. *Inc., Hoboken, New Jersey* 2006: 129.

26. Lin TY, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer. ; 2014: 740–755.

27. Andriluka M, Pishchulin L, Gehler P, Schiele B. 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. ; 2014: 3686–3693.

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

29. Yuan Y, Rao F, Lang H, et al. Hrformer: High-resolution transformer for dense prediction. arxiv 2021. *arXiv preprint arXiv:2110.09408*; 19.

30. Jiang T, Lu P, Zhang L, et al. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399* 2023.

31. Yukihiko A. USB Camera Motion Capture ThreeDPoseTracker Description. 2022.

32. Live3D . VTuber. 2023.

33. Jiang L, Cai L, Wu W, Zhou Z. Mirror world: creating digital twins of the space and persons from video streamings. *The Visual Computer* 2023.

34. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology* 1993; 3(3): 203–220.

6 | AUTHOR BIOGRAPHY

Ling Jiang received the M.S. degree from Ocean University of China, Qingdao, China, in 2018. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. Her main research interests include human pose estimation and 3D avatar.

Yuan Xiong received the B.S. degree from Beihang University in 2010 and the M.S. degree in computer science from Clemson University in 2014. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. His research interest includes camera calibration, multiple view geometry in computer vision and virtual reality.

Qianqian Wang received the B.S. degree from the Yanshan University in 2022. She is currently pursuing the master's degree with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China. Her research interest includes computer graphics and localization.

Tong Chen received the M.S. degree from North China Electric Power University, Beijing, China, in 2022. He is currently pursuing the Ph.D.degree with the School of Computer Science, Beihang University, Beijing, China. His research interests concern on computer graphics, virtual reality, 3D Reconstruction and localization.

Wei Wu received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1995. He is currently a Professor with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. He is the Chair of the Technical Committee on Virtual Reality and Visualization, China Computer Federation. His current research interests include virtual reality, wireless networking, and distributed interactive systems.

Zhong Zhou (Member, IEEE) received the B.S. degree from Nanjing University in 1999 and the Ph.D. degree from Beihang University, Beijing, China, in 2005. He is currently a Professor and the Ph.D. Adviser with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. His main research interests include virtual reality/augmented reality/mixed reality, computer vision, and artificial intelligence.

How to cite this article: Ling Jiang, Yuan Xiong, Qianqian Wang, Tong Chen, Wei Wu, and Zhong Zhou (2024), SADNet: Generating Immersive Virtual Reality Avatars by Real-time Monocular Pose Estimation, *Comput Anim Virtual Worlds*.