Supplementary Material for Cross-Camera Human Motion Transfer by Time Series Analysis

Paper ID #1121

Submitted to IEEE ICASSP'2024

EXPERIMENT

A. Qualitative Evaluations

In order to validate the effectiveness of our proposed algorithm, we apply it to real-world data procured from a dual-camera system [3]. By using a limited number of high-resolution video frames in conjunction with a complete series of low-resolution video frames, we initially carry out pose estimation using the OpenPose method [7]. This enables us to obtain a sequence of pose values.

As depicted by the green curves in Figure 1, the high-resolution video's estimated pose sequence is of superior quality. However, the pose sequence extracted from the low-resolution video is noticeably degraded by severe noise. Our motion transfer approach successfully mitigates this issue by refining the low-resolution pose values, as demonstrated in Figure 1(b). Our results effectively preserve the long-term correlation in the low-resolution pose sequence while successfully enhancing the quality of the low-resolution data through the transfer of motion patterns from the high-resolution sequence.

Following the procedures established in [3], we employ the SMPL model [1] to represent 3D bodies and utilize the HMR methodology [2] to reconstruct 3D human meshes from 2D video frames. While we can achieve satisfying results in straightforward cases (for example, where leg movement is minimal and arms are close to knees), as shown in Figure 2(a), it is clear that HMR struggles to accurately estimate 3D meshes in instances where the subject is walking at a brisk pace and swinging their arms vigorously, as depicted in Figure 2(b).

To enhance the quality of 3D human mesh reconstruction and to further underscore the efficacy of our motion transfer algorithm, we apply our method to high-resolution and low-resolution real-world videos of a pedestrian walking on a street, captured by our dual-camera system. As evidenced in the right of Figure 3, the 3D human meshes reconstructed without our method are inaccurate and implausible, with obvious anomalies such as unnaturally extended arms resulting from the inferior quality of the low-resolution video. However, the use of our motion transfer method results in the accurate correction of these human poses.

Taking our evaluation of the motion transfer method a step further, we integrate it into a downstream vision task. The goal is to embed high-resolution human details into low-resolution videos. As depicted at the top of Figure 4, without our algorithm, the direct synthesis of human details appears inaccurate. For example, obvious jittering can be observed due to the blurred frames, especially in the head and leg regions, with the feet assuming extremely unnatural poses. Conversely, as demonstrated in the bottom row of Figure 4, the application of motion transfer significantly reduces motion jittering, highlighting the utility of our approach.

B. Quantitative Evaluations

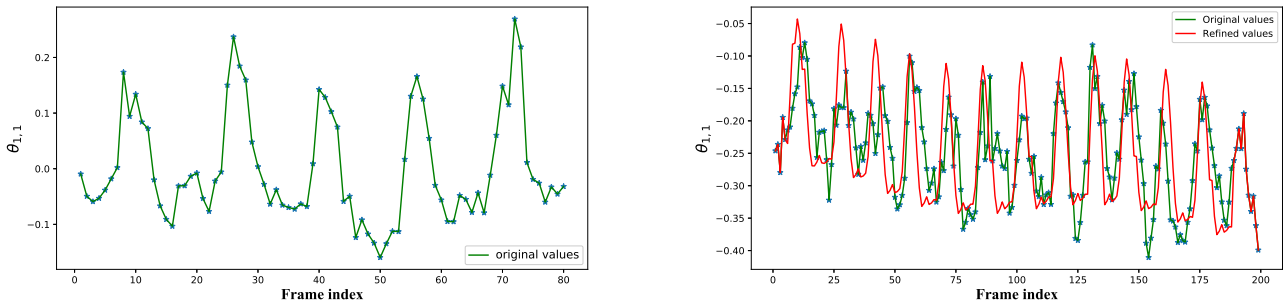
Given that our motion transfer approach is applied to the coordinate points of human body joints derived from real-world scenarios, there are no ground truths available, making it challenging to quantify the performance of the motion transfer directly. To address this, we resort to a classic downstream task — video super-resolution through texture transfer — to quantitatively evaluate the proposed motion transfer method.

Specifically, we establish synthetic data sets derived from the MPII dataset [4], which comprises a considerable volume of image sequences centered around human figures and exhibiting a range of human poses. Most MPII sequences consist of about 40 frames. We designate a single frame as the high-resolution frame, while the remaining frames are downsampled by a factor of 8 and treated as low-resolution frames. Following this, we recover the high-resolution details for the downsampled video with the aid of the proposed motion transfer method. Further details regarding the implementation can be found in [3].

As demonstrated in Fig. 5 [3], we conduct a comparative analysis of our method with the methods proposed by Wang et al. [5] and Tao et al. [6]. In the context of synthetic data, as shown in Fig.5, the approaches in [5] and [6] result in blurry outcomes. In contrast, our method generates visually pleasing results, despite lower PSNR values, especially around the facial region.

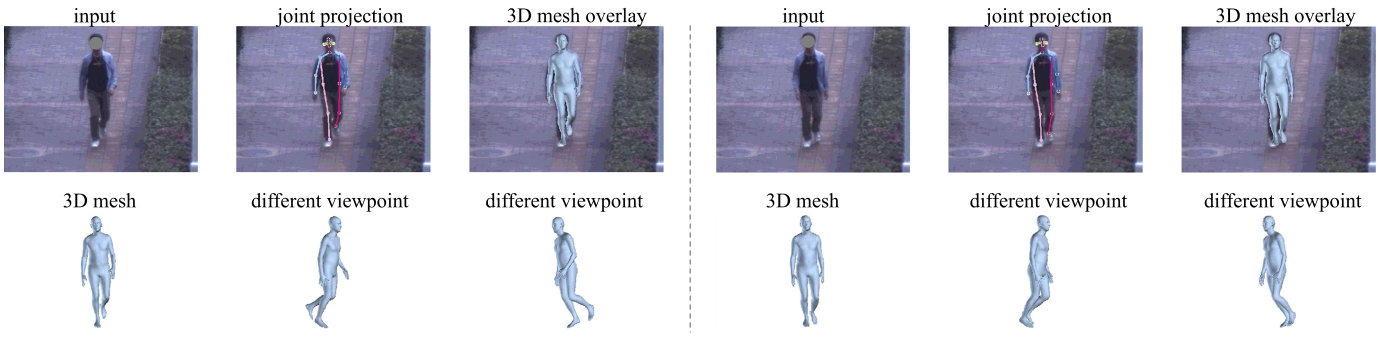
REFERENCES

[1] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015.

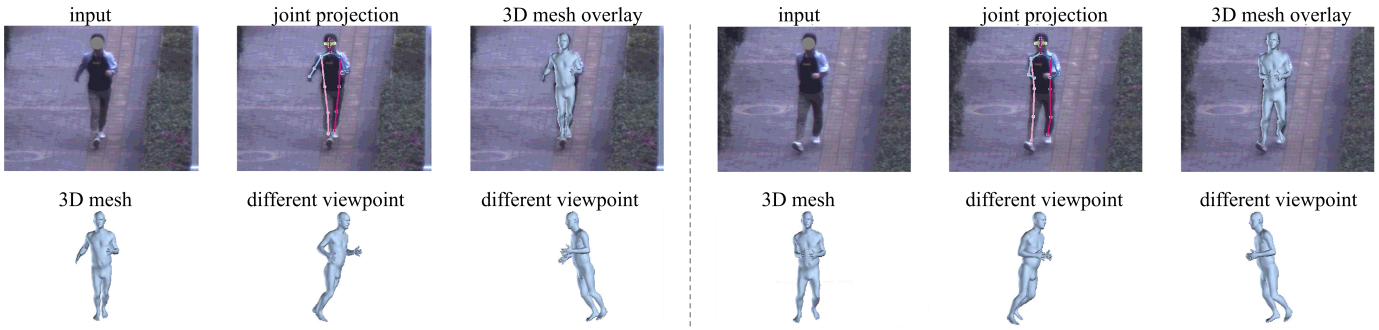


(a) (b)

Fig. 1: HR and LR motion data. (a) and (b) represent the $\theta_{1,1}$ value of the HR and LR motion data, respectively.



(a)



(b)

Fig. 2: The SMPL model [1] is adopted to represent 3D body and the HMR [2] is performed to reconstruct 3D human meshes from 2D video frames. (a) and (b) show successful cases and failure cases, respectively. Since we collect and use real-life data in this figure, the human faces are masked out to protect personal privacy.

[2] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018.

[3] Guanghan Li, Yaping Zhao, Mengqi Ji, Xiaoyun Yuan, and Lu Fang. Zoom in to the details of human-centric videos. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3089–3093. IEEE, 2020.

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[5] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[6] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017.

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

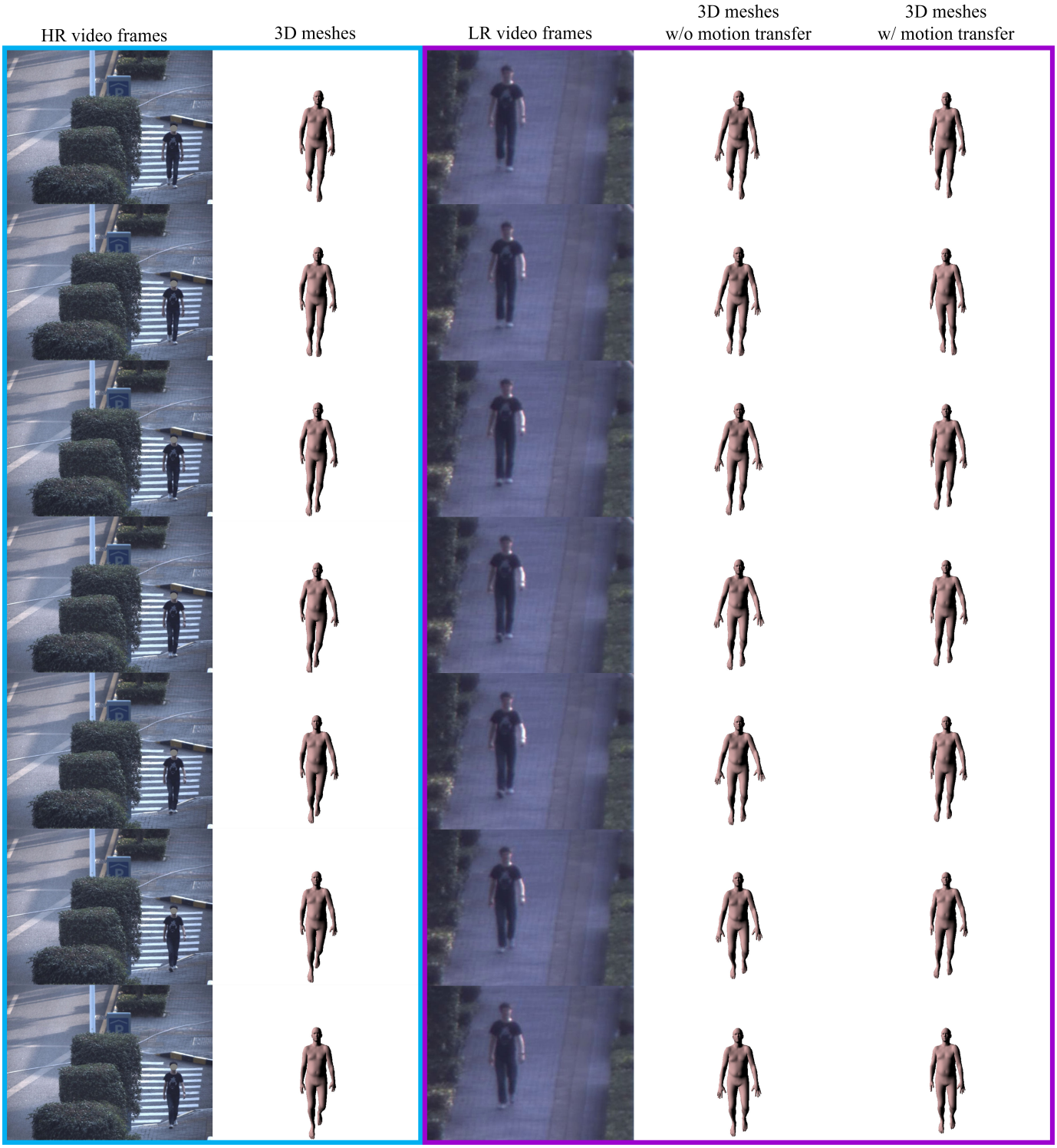


Fig. 3: Performance on 3D human mesh reconstruction using our motion transfer. We observe the wield opening arms due to the low quality of the LR video. In contrast, the results with motion transfer substantially correct the human poses. Since we collect and use real-life data in this figure, the human faces are masked out to protect personal privacy.

Fig. 4: Performance on human detail synthesis across heterogeneous cameras. Top: the synthesized results without our motion transfer method. Bottom: the synthesized results with our method. Without motion transfer, there are intensive jitters on the head and legs. Since we collect and use real-life data in this figure, the human faces are masked out to protect personal privacy.

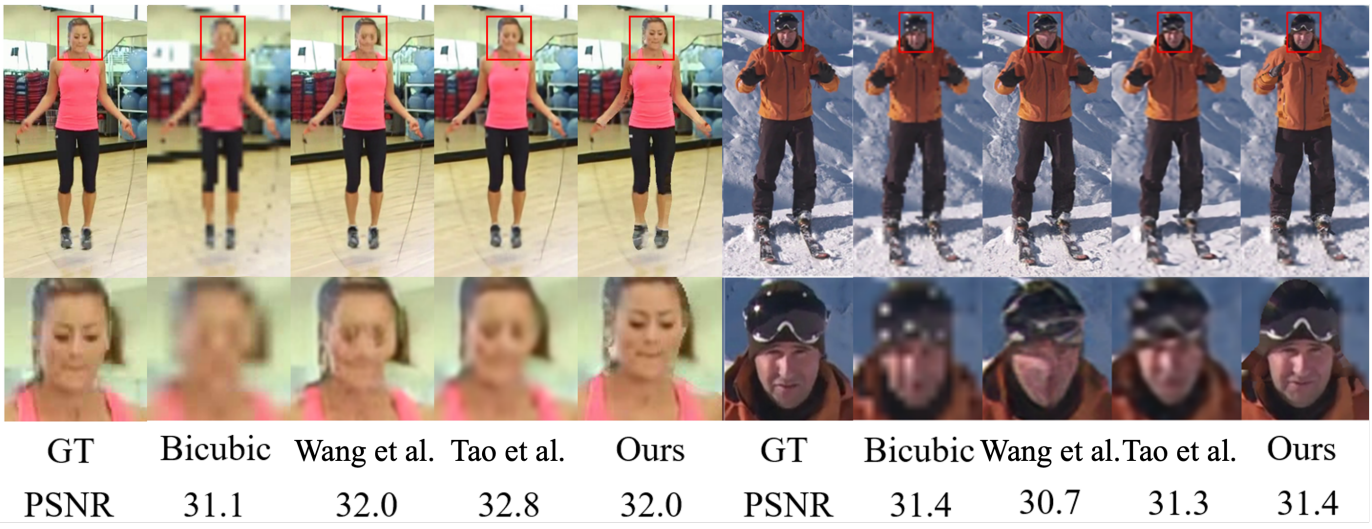


Fig. 5: Results on $\times$ 8 RefSR [3] in the synthesised MPII dataset [4]. We benchmark our results against established methods, including Bicubic interpolation and the algorithms proposed by Wang et al. [5], and Tao et al. [6]. Although our method manifests a marginally lower peak signal-to-noise ratio (PSNR), it consistently delivers superior visual quality of the human body in comparison with the other tested methodologies. This underscores the strengths of our approach in maintaining high visual fidelity despite lower quantitative metrics, ultimately emphasizing its potential effectiveness for applications in the domain of computer vision that demand high-quality rendering of human figures.