Coordinate Transformer: Achieving Single-stage Multi-person Mesh Recovery from Videos — Supplementary Materials

1. Introduction

Here we provide more implementation details of the training and evaluation; additional evaluation results on in-the-wild datasets; additional ablation studies with visualizations; and additional visualization results on internet videos. Further, we perform a comparisons to state-of-the-art methods on extreme scenarios to complement our comprehensive analysis. Finally, a video demo is provided in our supplementary material to show additional qualitative video results.

2. Implementation Details

Similar to [12], we resize the image sequences to a size of 512×512. The size of the backbone features is $H_f = W_f = 128$, and the maximum number of detections N is set to 64. The time window size T is set to 8. The spatial loss weights are set to w_{cm} = 160, w_{mpj} = 360, w_{pmpj} = 400, w_{pj2d} = 420, w_{pose} = 80, w_{shape} = 1, w_{prior} = 1.6, and are fixed in all training steps. The temporal loss weights are set to w_{accel} = 200, w_{aj3d} = 300, w_{sm} = 100, and are only used when fine-tuning on the video datasets. The threshold t_c of the Body Center Heatmap is set to 0.2. The learning rate lr of the Adam optimizer is set to $5\times\ 10^{-5}$ and the batch size B is 16. Training is performed in an end-to-end manner directly from image or video inputs.

2.1. Training Datasets

Since CoordFormer is a novel approach for multi-person videos, our training and evaluation focuses on the most relevant dataset (3DPW [14]), while other datasets are used as supplements to improve generalization and enhance prediction accuracy. For completeness, we have included the dataset details below.

3DPW [14] is a challenging outdoor dataset with more than 51,000 frames of 7 actors in various clothing styles. The dataset includes numerous frames with multi-person interactions and all the raw ground-truth markers are recorded via Inertial Measurement Units, which provide accurate ground truth annotations.

Human3.6M [2] is an indoor, multi-view, single-person 3D human pose estimation dataset. The extended SMPL model annotations are generated from sparse marker data. Following [12], we use 5 subjects (S1,S5,S6,S7,S8) for training.

MPI-INF-3DHP [10] is an indoor, multi-view, single-person 3D human pose dataset with some noise.

MuCo-3DHP [10] is an extended version of MPI-INF-3DHP using data augmentation. The authors replace the background with real-world images and place 1 to 4 subjects on the background to facilitate a range of inter-person overlap and activity scenarios.

In-the-wild 2D datasets MPII [1] and LSP [3, 4] are in-the-wild 2D datasets, which are collected using Amazon Mechanical Turk. Annotation quality of the 2D labels is improved by modelling the annotator error using iterative procedures.

2.2. Training Steps

Existing video-based methods use an explicit 2D detector and a tracker to model the temporal relationship of a particular individual. Instead, CoordFormer employs Body Center Attention as an implicit detector and the Spatial-Temporal Transformer to learn temporal relations. This allows CoordFormer to not only leverage video data, which often can be restricted in the multi-person setting, but also leverage available image datasets. This can be facilitated by training CoordFormer in three steps: First the spatial branch of CoordFormer will be trained like most existing single image-based methods [12, 8], while the second step consists of fine-tuning the spatial and temporal branch on the video dataset without 3DPW. Finally, CoordFormer is fine-tuned with the 3DPW dataset in the third step.

2.3. Training Dataset Ratio

To obtain the best results, we follow EFT [6] and SPIN [9] to batch data according to the dataset sample ratios. To train the spatial branch of CoordFormer without using temporal information, we incorporate 30% MPI-INF-3DHP, 10% LSP, 15% MPII, 20% MuCo-3DHP and 25% Human3.6 M into training in the first step. Next, we fine-tune the model on the Human3.6 M dataset in the second step. Finally, the model is fine-tuned on 3DPW to achieve the final best performing model.

2.4. Evaluation Strategy

For a fair comparison, we use CoordFormer after the second training-step for evaluation following *Protocol 1 and 2*, and use the fine-tuned model after the third training-step to evaluate the best results (*Protocol 3*).

Table 1: Comparsions resuls on CMU Panoptic[5] and MuPoTs[11] according to PAMPJPE metric.

Methods	**CMU Panoptic**					**MuPoTs**
	Haggling	**Mafia**	**Ultim**	**Pizza**	**Mean**	
ROMP	68.16	79.25	**76.89**	85.25	77.39	93.00
CoordFormer	**66.82**	**77.23**	77.83	**83.03**	**76.23**	**88.02**

3. Further evaluation results on in-the-wild datasets.

According to the PAMPJPE metric, the comparison results on CMU Panoptic[5] and MuPoTs[11] are reported for comprehensive evaluations under in-the-wild multi-person scenarios. All the methods are directly evaluated without any fine-tuning. As shown in Tab. 1, CoordFormer outperforms ROMP in almost all activities. Moreover, as shown in Fig. 1, CoordFormer performs better detection and pose estimation.

4. Further Ablation Study With Visualization

To further show the effectiveness of BCA and CAA, we collect the evaluation results on the 3DPW validation set during the training and compare the performance of the ablation models. All the models are trained for 10 epochs, following the same setting as the first training step in Sec. 2.2. For the sake of readability, the notations of different model settings are summarized in Tab. 2.

How BCA accelerates the training process. As shown in Fig. 2b and Fig. 2c, models with BCA mechanism achieve better performance and more stable convergence under different model settings. More specially, CF_{bca} outperforms CF_{None} by 8.2% and 6.8% according to the MPJPE and PAMPJPE metrics, while S-CF_{None} could not converge for the same training datasets.

Table 2: The notations of CoordFormer under different settings. Note, here ”splitting” refers to adopting the tokenization method that splits the features into patches and extract tokens from them and CAA is not used to highlight the effectiveness of BCA.

	w/o BCA	w/ BCA
splitting	S-CF_{None}	S-CF_{bca}
not splitting	CF_{None}	CF_{bca}

Table 3: Ablation study of CAA under different training steps.

Steps	Methods	MPJPE↓	PAMPJPE↓	PVE↓
Step 1	CF_{bca}	**101.73**	**55.68**	**117.91**
	CoordFormer	103.95	58.03	120.67
Step 2	CF_{bca}	97.14	56.01	112.69
	CoordFormer	**95.27**	**54.58**	**110.35**
Step 3	CF_{bca}	83.19	50.62	99.21
	CoordFormer	**79.41**	**46.58**	**94.44**

How CAA preserves pixel-level representations. Patch-level tokenization of standard vision transformers leads to feature disorder and feature partition, which occurs when one patch contains multiple person and when the body center is located on the boundary line of the patch, respectively. The design of CAA allow us to keep pixel-level representations and avoid the spatial information degradation. As shown in Fig. 2d, S-CF_{None} results in extensive fluctuations in performance and crashes halfway through training due to sudden excessive losses. Moreover, we observe that CF_{None} converges quicker than S-CF_{bca}.

However, it is not enough to build spatial-temporal constraints only based on pixel-level tokenization. Tab. 3 illustrates CAA's effectiveness to capture coordinate information across frames. Although CF_{bca} performs better for single-image regression, CoordFormer demonstrates superior modeling of spatial-temporal relations.

Qualitative ablation study of visualization comparison. Fig. 3 illustrates that 1) BCA and CAA have to be combined to facilitate accurate Body Center heatmap prediction using the CoordFormer, 2) BCA can enhance the confidence on the body center, especially under person-occlusion scenarios, 3) CoordFormer w/o CAA regresses the mesh only based on BCA, which improves results on certain individuals, but fails to model temporal relations, thus degrading pose and shape coherence.

Whether BCA and CAA conform to our assumptions. As shown in Tab. 4, CoordFormer with CAA improves performance by learning temporal information in training step 2 according to both MPJPE and PAMPJPE, while CoordFormer with only BCA obtains worse PAMPJPE. This illustrates that BCA focuses on single-image and CAA focuses on temporal information, which is consistent with our implicit detection and tracking assumptions. Moreover, as shown in Fig. 1, CoordFormer performs better detection and pose estimation.

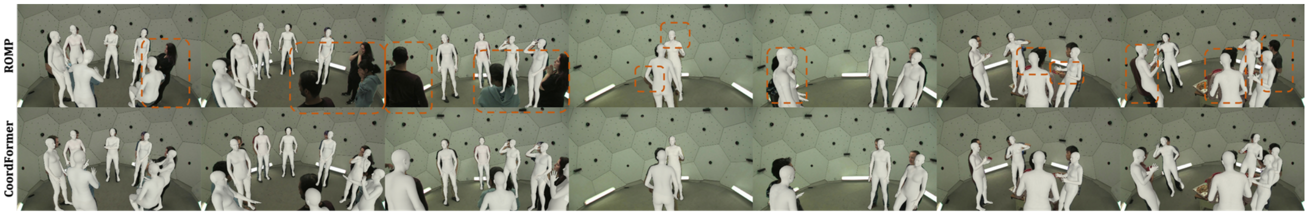


Figure 1: Partial visualization comparison between CoordFormer and ROMP on CMU Panoptic[5] dataset.

Table 4: Ablation study under 3DPW.

Methods		MPJPE ↓			PAMPJPE ↓			PVE ↓		
BCA	CAA	step1	step2	step3	step1	step2	step3	step1	step2	step3
✓		**101.73**	97.14	83.19	**55.68**	56.01	50.62	**117.91**	112.69	99.20
	✓	103.61	99.94	82.20	57.64	55.63	48.84	120.04	114.82	98.23
✓	✓	103.95	**95.27**	**79.41**	58.03	**54.58**	**46.58**	120.67	**110.35**	**94.44**

5. Further visualization results on the internet videos.

We further test CoordFormer on internet videos, especially sports videos. As shown in Fig. 4, CoordFormer effectively obtains the multi-person mesh from a variety of videos.

6. Further Visualization Compared to State-of-the-art Methods.

To show the superior performance of CoordFormer beyond simple in-the-wild scenarios, we compare CoordFormer to the best pre-trained ROMP [12] and BEV [13] models. Note, these were trained on a considerably larger number of datasets and for BEV, leverage a larger Body Center heatmap with a size of 128, resulting in more capacity to provide accurate and precise predictions. While an unfair comparison from CoordFormer's perspective, we observe that CoordFormer still obtains preferable results.

Qualitative results on internet videos with small targets. Given the precise coordinate information to refine the Body Center heatmap, CoordFormer is able to better detect people in the video, especially the small targets, which is crucial for 3D human mesh recovery from athletic sports videos and aerial videos. As shown in Fig. 5, CoordFormer obtains great detection results and achieves the best visualization results for small targets. While CoordFormer is not able to provide mesh results for all people, CoordFormer achieves significant improvements on the accuracy of the Body Center heatmap and Camera map compared to ROMP and BEV.

Qualitative results on internet videos with low resolution. As shown in Fig. 6, CoordFormer displays superior robustness to videos with different resolution. Specifically, CoordFormer can still maintain its performance even for videos with low resolution of 64×36. Compared with state-of-the-art video-based [7, 15] methods that are equipped with 2D detectors, Fig. 7 illustrates that CoordFormer achieves the best results over methods with requiring explicit detection.

7. Social Impact

While there are a wide range of application domains where video-based 3D human mesh recovery will be beneficial, such as for instance in physical therapy and virtual reality, there are also potentially negative application scenarios. For instance, these approaches could be used in malicious contexts to obtain a large amount of private body data or for surveillance purposes. Consequently, CoordFormer is released as a research tool only.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693. IEEE, 2014. 1

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, pages 1325–1339, 2013. 1

[3] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, page 5. Citeseer, 2010. 1

[4] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472. IEEE, 2011. 1

[5] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, pages 3334–3342, 2015. 2, 3

[6] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *ECCV*, pages 68–84. IEEE, 2020. 1

[7] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263. IEEE, 2020. 3, 9

[8] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137. IEEE, 2021. 1

[9] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261. IEEE, 2019. 1

[10] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516. IEEE, 2017. 1

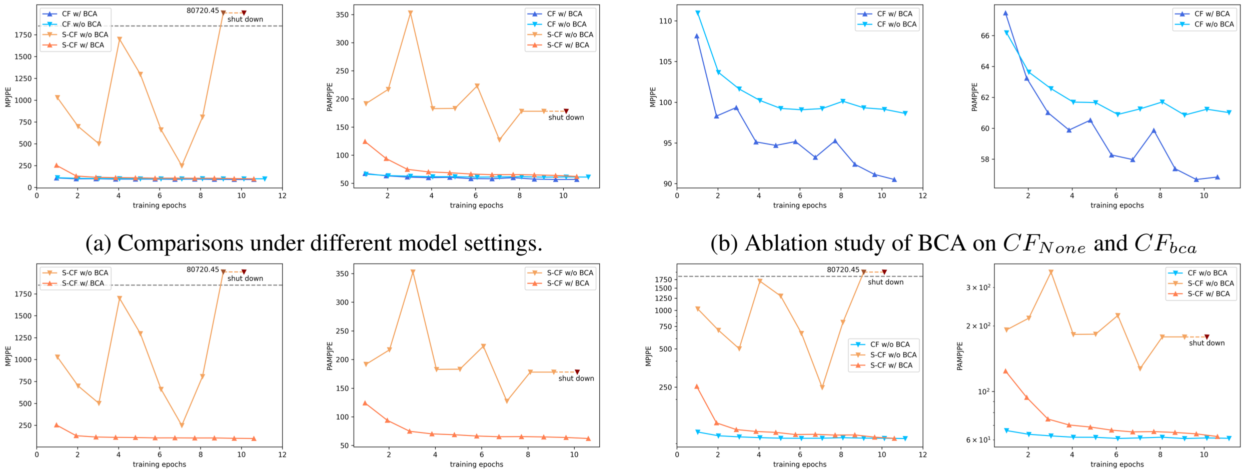
[11] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, pages 120–130. IEEE, 2018. 2

[12] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188. IEEE, 2021. 1, 3, 7, 8, 9

[13] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252. IEEE, 2022. 3, 7, 8, 9

[14] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617. Springer, 2018. 1

[15] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *CVPR*, pages 13211–13220. IEEE, 2022. 3, 9



(a) Comparisons under different model settings.

(b) Ablation study of BCA on CF_{None} and CF_{bca}

(c) Ablation study of BCA on $S\text{-}CF_{None}$ and $S\text{-}CF_{bca}$

(d) Comparison of different methods for obtaining tokens.

Figure 2: Further ablation study of BCA and CAA at the first training step.

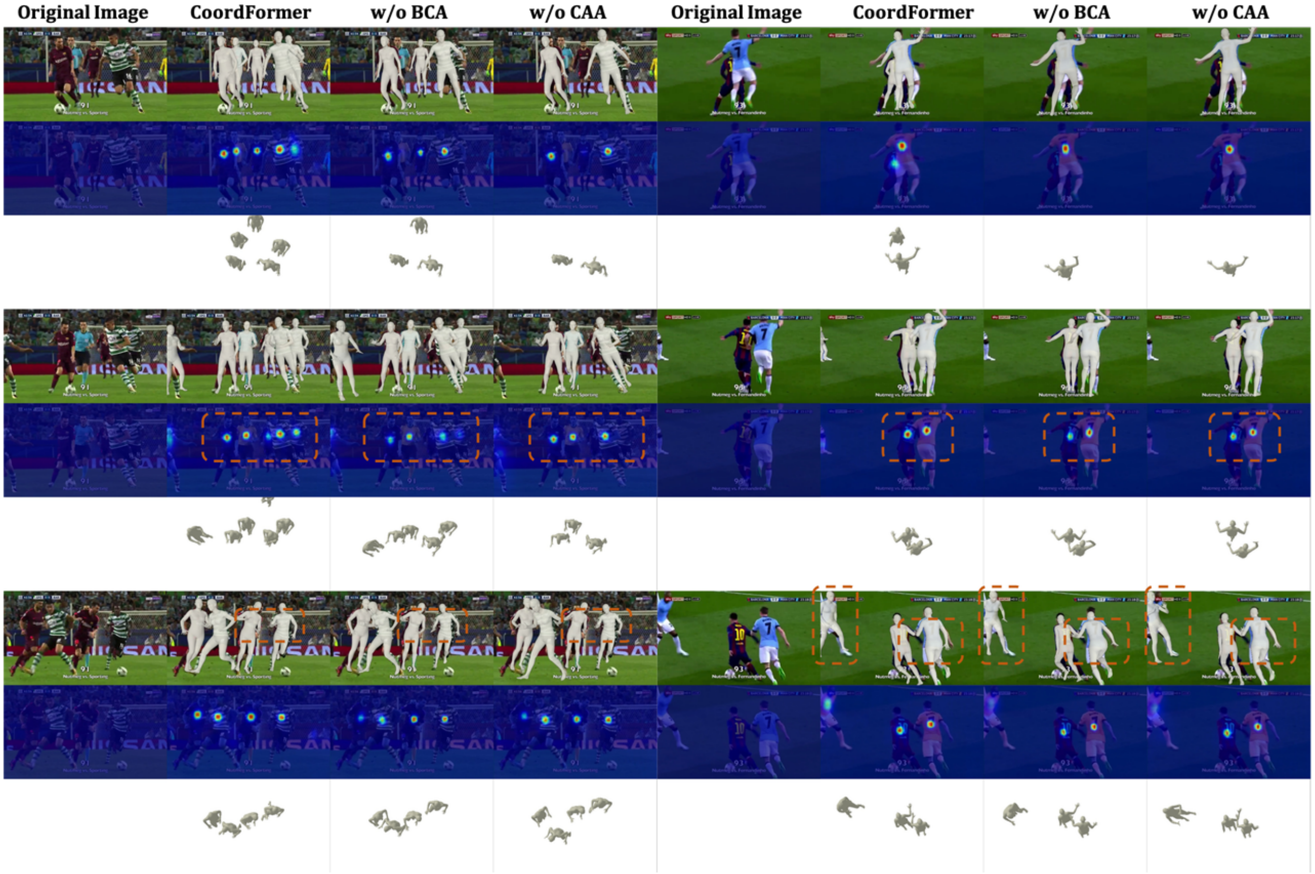


Figure 3: Further ablation study of visualization comparison.

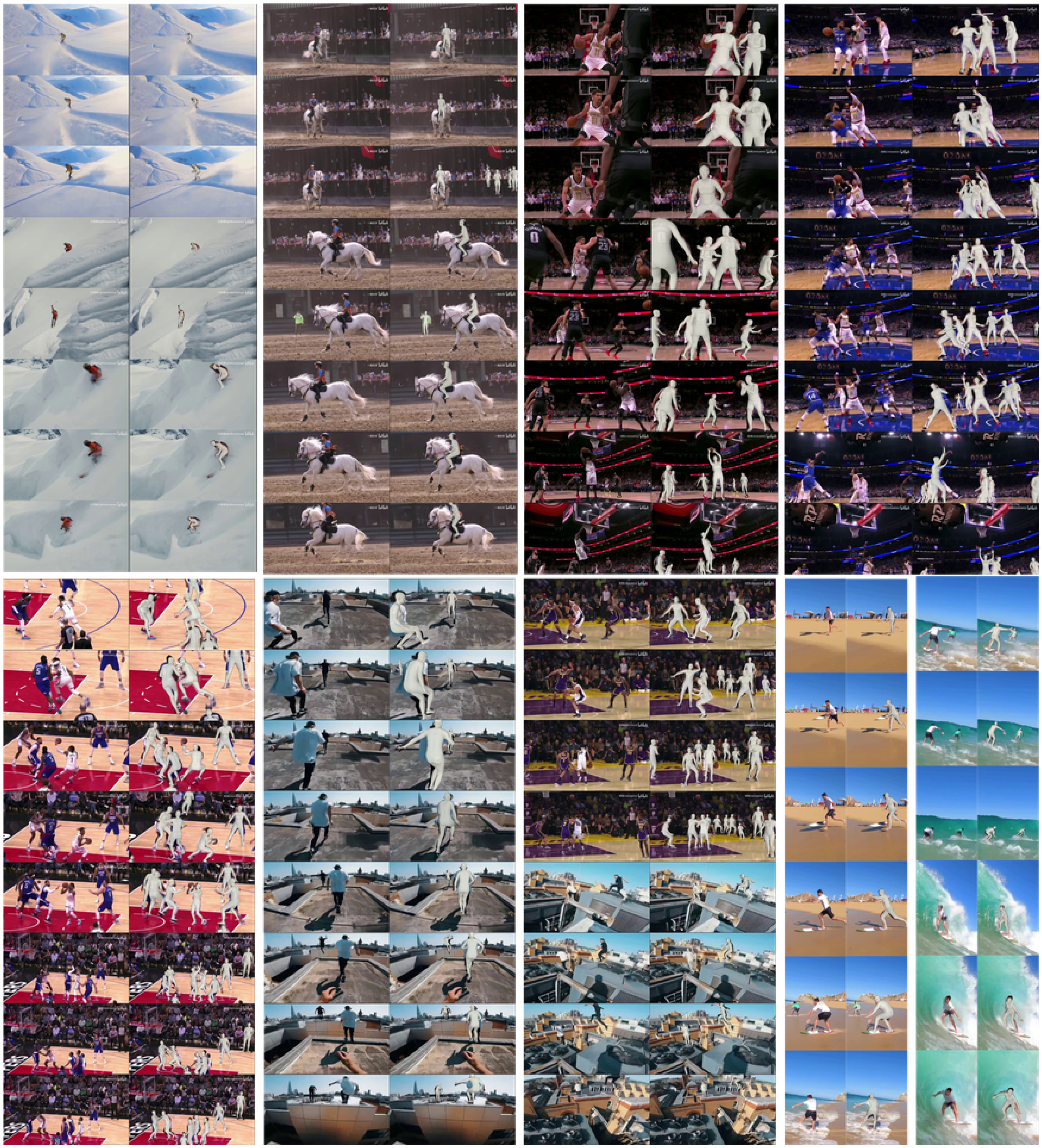


Figure 4: Further visualization results of CoordFormer on the internet videos.

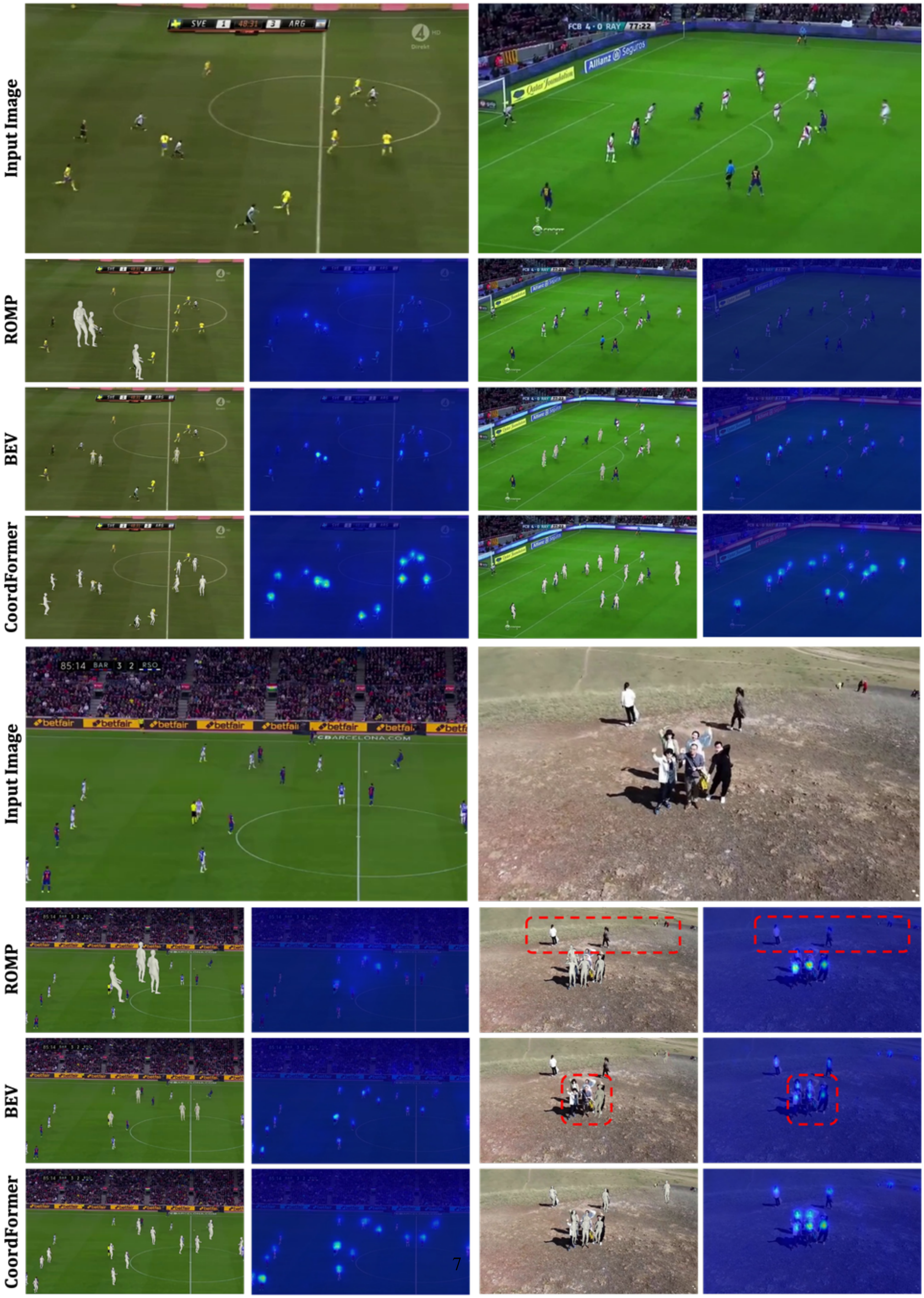


Figure 5: Qualitative results of ROMP [12], BEV [13] and CoordFormer on the internet videos with small targets.

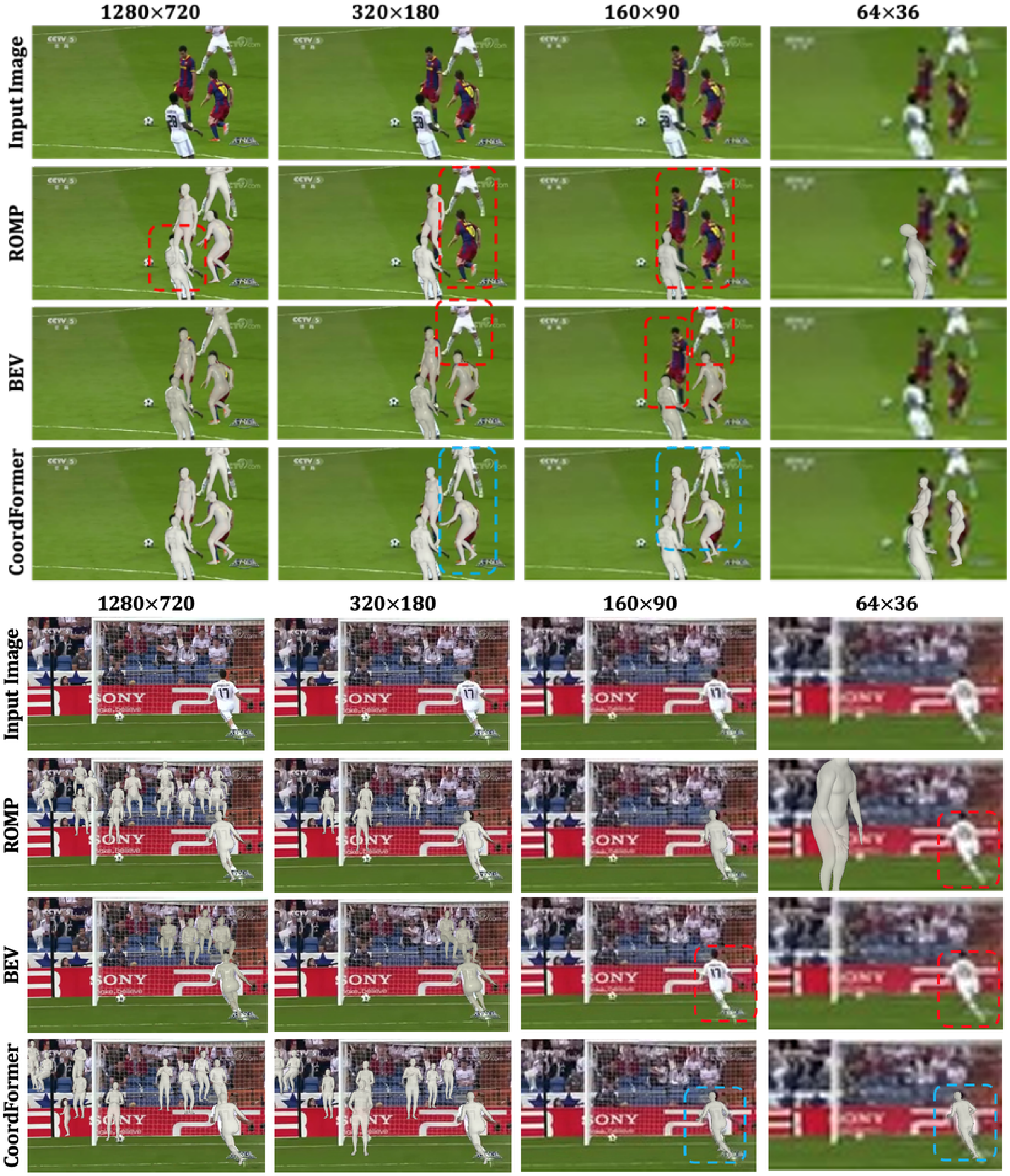


Figure 6: Qualitative results of ROMP [12], BEV [13] and CoordFormer on the internet videos with low resolution.

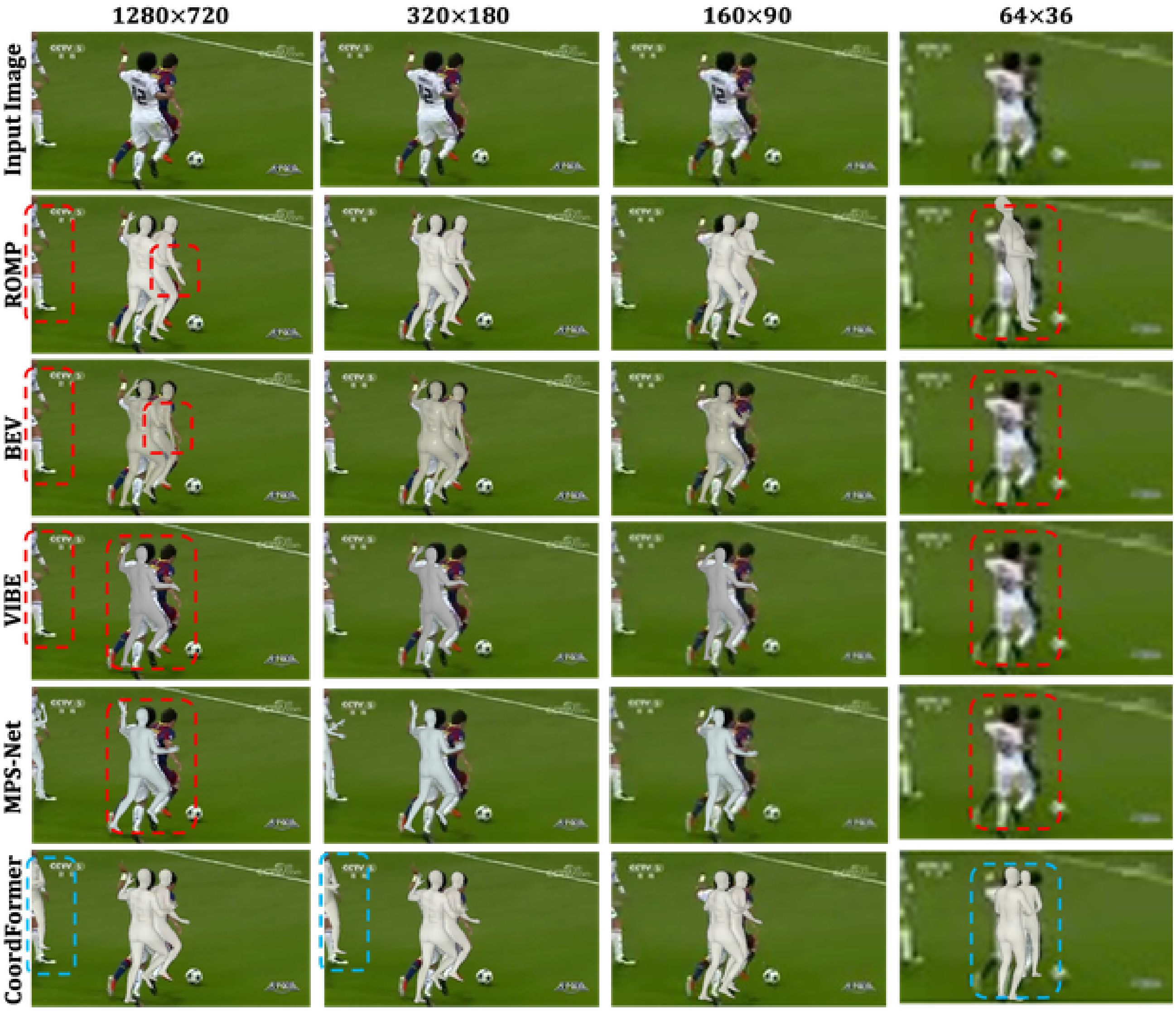


Figure 7: Qualitative results of ROMP [12], BEV [13], VIBE [7], MPS-Net [15] and CoordFormer on the internet videos with low resolution.