Appendix

A. Off-the-shelf Models

A.1. 2D Pose Estimation

As mentioned in Sec. 6.1, we adopt a pretrained HRNet [32] model for 2D pose estimation. Since the model was originally trained on MSCOCO [21], which uses slightly different format of human body keypoints, we fine-tune the model to convert the keypoints to the desired format. (Note that MSCOCO, MuPoTS-3D, and CMU-Panoptic use different keypoints, while MuPoTS-3D, MuCo-3DHP, and MPI-INF-3DHP share the same one.) The model is fine-tuned on the MuCo-3DHP dataset for MuPoTS-3D testing, and fine-tuned on CMU-Panoptic training dataset for CMU-Panoptic testing, respectively. Each of which is fine-tuned for 20 epochs with a learning late of 10^{-4}, which is 10 times smaller compared to the original learning rate at training.

A.2. 2D Pose Tracking

In Sec. 6.1, we mention that we use ByteTrack [42] for re-identification of each individual. We also merge the appearance gallery idea [37] to consider appearance variation caused by movements. While tracking individuals frame by frame, the most recent 100 appearance features are stored in their tracklet. For measuring similarities, in total three similarities are used. In addition to the appearance feature and IoU, we also consider the similarity between poses as well.

We observe clean and accurate tracking is important for end-to-end performance. Specifically, performance of our model is quite sensitive to the tracking result, especially, to the weights among 3 similarities above. Giving a higher weight to the appearance similarity might help the model to match re-appearing individuals after a period of heavy-occlusion, meanwhile it does not consider about their location. In contrast, giving a higher weight to IoU or pose similarity might help the model to accurately track the motion dynamics of individuals, but as a trade-off, it confuses the model to match re-appearing individuals. We empirically find the optimal weights and use $\{0.4, 0.3, 0.3\}$ for each appearance, IoU, and pose similarity, respectively.

One thing to note is that the end-to-end evaluation metrics are not significantly affected by a few tracking failures, as it usually match each predicted individual with the ground truth frame-by-frame. For a real-world application, largely disturbing tracklets might be omitted and only some clean tracklets could be chosen for the sake of reliability.

B. Details on Data Augmentation

The data augmentation hyperparameters, $\alpha, \beta, \gamma, \theta, \varphi$, are empirically chosen, considering typical movement range of a person for translations and approximate angular distance between cameras of the source dataset for rotations, respectively. α, β are sampled from a Gaussian $\mathcal{N}(0, 6.0^2)$ and γ is randomly chosen among $\{-1.0, 0, 1.5, 3.0\}$, where the unit is meters. θ is uniformly sampled within $[-\pi/4, \pi/4]$, and φ is randomly chosen among $\{-\pi/6, 0, \pi/6\}$.

Validity of a Training Example. Once a training example is generated, we check its validity. First of all, the depth z of all target 3D keypoints should be positive. Otherwise, a subject with negative depth will appear flipped both vertically and horizontally after projection.

Also, as we constrain the number of people appearing in the scene to be consistent throughout the temporal receptive field (*i.e.* people are assumed not to be jumping in or fading out), we force the resulting trajectory to be entirely located within the 2D frames. Precisely, we keep the root key points to appear within the image boundary but let other joints potentially be out of the scene. For this, we might naively filter out examples that violate the constraints and regenerate, but this is not efficient. Instead, we apply PR, GPT, and GPR first, and PT at the last. Unlike other operations, we can constrain the feasible range for PT individually, satisfied simply by solving a constrained linear programming:

$$\begin{aligned} 0 \le f_u \frac{x + \Delta x}{z + \Delta z} + c_u < W, \\ 0 \le f_v \frac{y + \Delta y}{z + \Delta z} + c_v < H, \\ 0 \le z + \Delta z, \end{aligned} \tag{4}$$

where (x, y, z) is an original root joint in the 3D space, $(\Delta x, \Delta y, \Delta z)$ is the amount of displacement applied to this subject, converted from (α, β) on the basis $\{\mathbf{b}_1, \mathbf{b}_2\}$ to the standard basis $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$, and W, H is the width and height of the image.

C. More Quantitative Results

Tab. I lists the performance of our model and baselines on individual test videos in MuPoTS-3D. We observe that our proposed method, POTR-3D, outperforms baselines on most videos, especially when severe occlusion occurs (TS 2, 13, 14, 18, 20).

For reference, Fig. I illustrates the conventional camera settings in CMU-Panoptic, and the ones we use in Sec. 6.3.

D. More In-the-wild Examples

Fig. II–III illustrate qualitative results on MuPoTS-3D of ours and baselines from side view, and top view. They appeal that the depth estimation of POTR-3D is more accurate and smooth.

Method	PCK_{rel}(%)↑	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20
SingleStage [14]	80.9	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SMAP [44]	73.5	88.8	71.2	77.4	77.7	80.6	49.9	86.6	51.3	70.3	89.2	72.3	81.7	63.6	44.8	79.7	86.9	81.0	75.2	73.6	67.2
SDMPPE [26]	81.8	**94.4**	77.5	79.0	81.9	85.3	72.8	81.9	**75.7**	**90.2**	**90.4**	79.2	79.9	75.1	72.7	81.1	89.9	**89.6**	81.8	81.7	76.2
POTR-3D (Ours)	**83.7**	92.0	**80.2**	**83.7**	**84.0**	**85.4**	**75.1**	**91.5**	74.3	70.7	88.4	**85.6**	**86.5**	**83.1**	**77.1**	**82.8**	**90.8**	86.8	**87.5**	**85.7**	**82.6**
Method	PCK_{abs}(%)↑	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS18	TS19	TS20
VirtualPose [31]	44.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SingleStage [14]	39.3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SMAP [44]	35.2	21.4	22.7	58.3	27.5	37.3	12.2	49.2	40.8	**53.1**	43.9	**43.2**	43.6	39.7	28.3	49.5	23.8	18.0	26.9	25.0	38.8
SDMPPE [26]	31.5	**59.5**	**44.7**	51.4	46.0	52.2	27.4	23.7	26.4	39.1	23.6	18.3	14.9	38.2	26.5	36.8	23.4	14.4	19.7	18.8	25.1
POTR-3D (Ours)	**50.9**	50.1	42.1	**71.0**	**60.5**	**58.6**	**50.4**	**66.9**	**41.5**	50.0	**69.6**	42.3	**49.2**	**63.2**	**49.3**	**69.0**	**35.6**	**36.9**	**35.3**	**29.3**	**46.3**

Table I. **Quantitative Comparison on MuPoTS-3D for Individual Test Videos.** The best scores are marked in boldface.

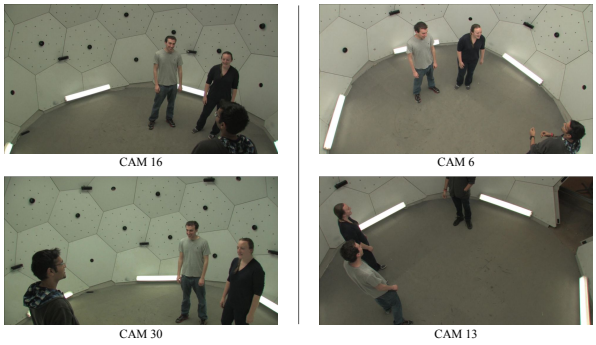

Figure I. **Camera view points of CMU-Panoptic.** (*Left*) Cameras used in conventional benchmark for both training and testing. (*Right*) Cameras used for our experiment in Sec. 6.3.

Fig. IV–XI illustrate more qualitative results of our model on several challenging in-the-wild videos. We present the results of 10 frames from the frontal view per video to demonstrate both accuracy and smoothness. It robustly operates even in highly challenging situations, such as heavy occlusions, dynamic motions, and non-static camera movement. The results from other views are provided in the demo video at https://www.youtube.com/@potr3d. To create this video, we use a POTR-3D model trained on the augmented dataset from MPI-INF-3DHP with Aug4, and assume a general focal length (*i.e.*, 1500) to denormalize the depths. At last, some examples in Fig. XII include additional failure cases, caused by tracking failure, and depth ambiguity.

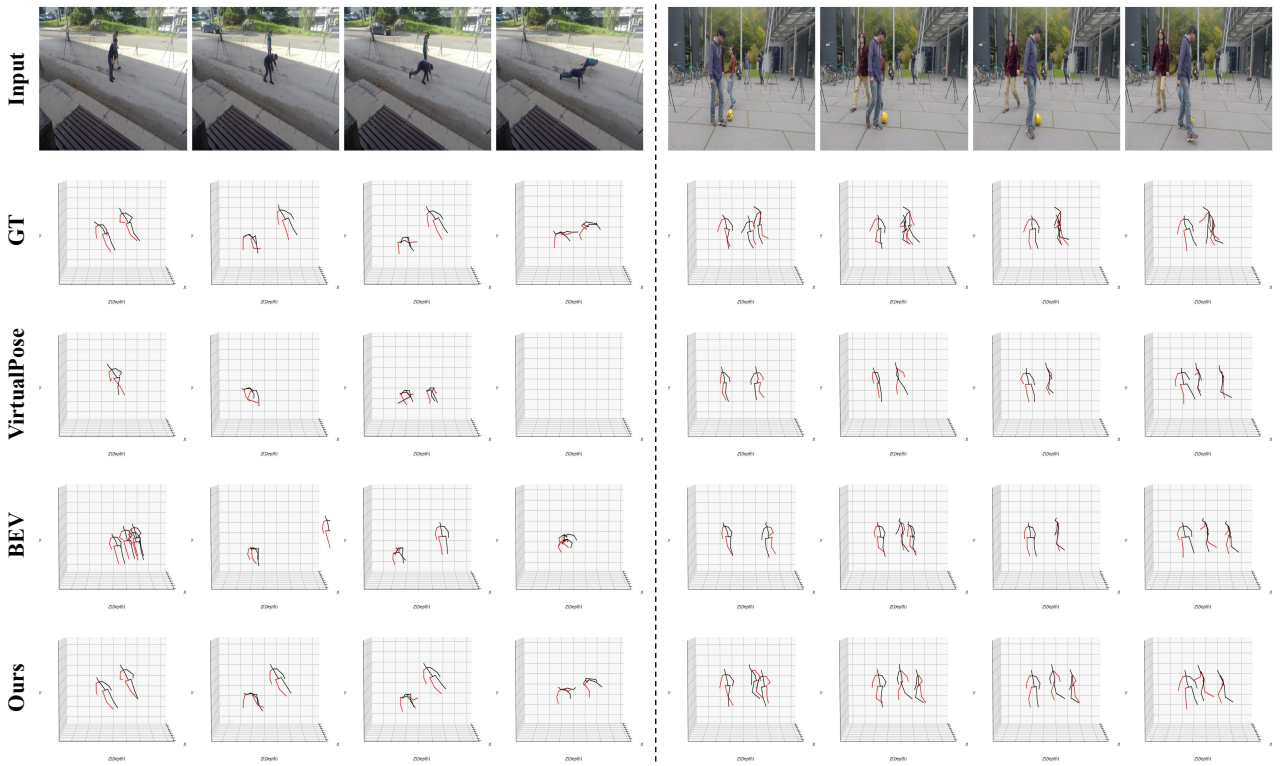

Figure II. **Qualitative Results on MuPoTS-3D** of ours and recent baselines **(Side View)**.

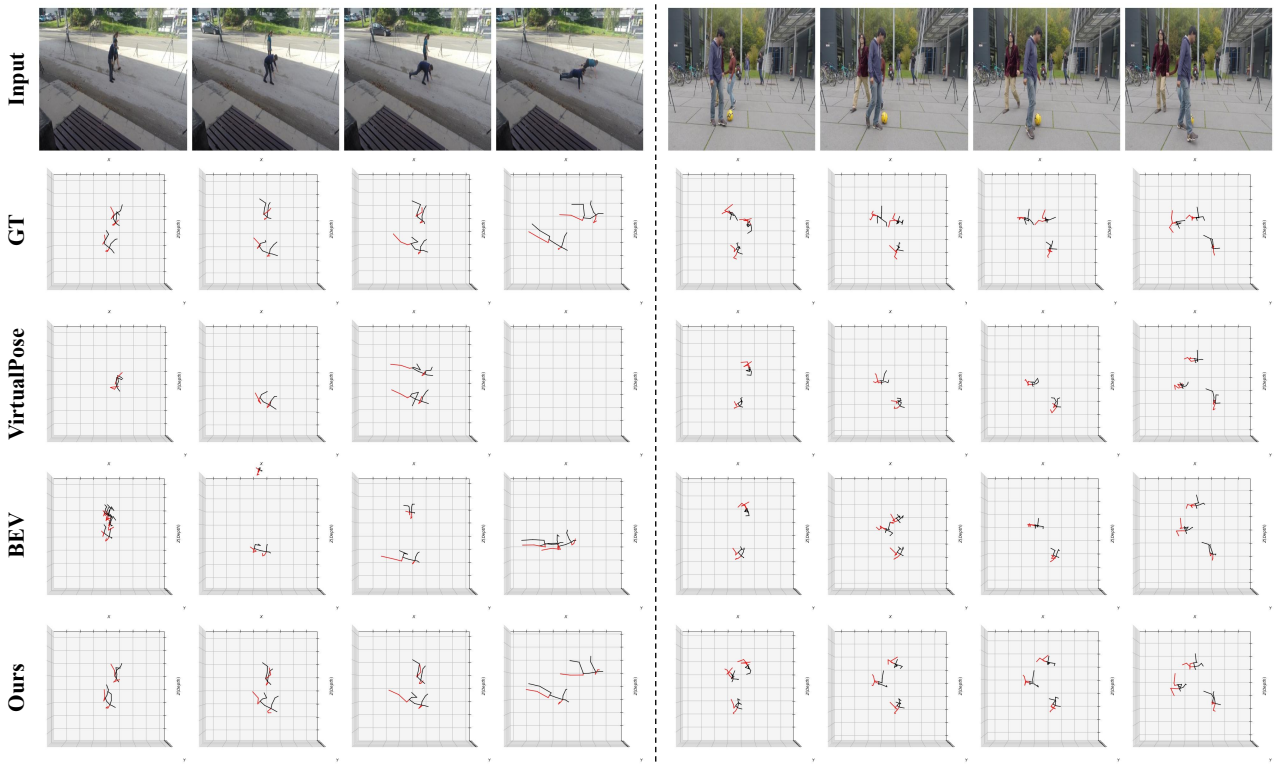

Figure III. **Qualitative Results on MuPoTS-3D** of ours and recent baselines **(Top View)**.

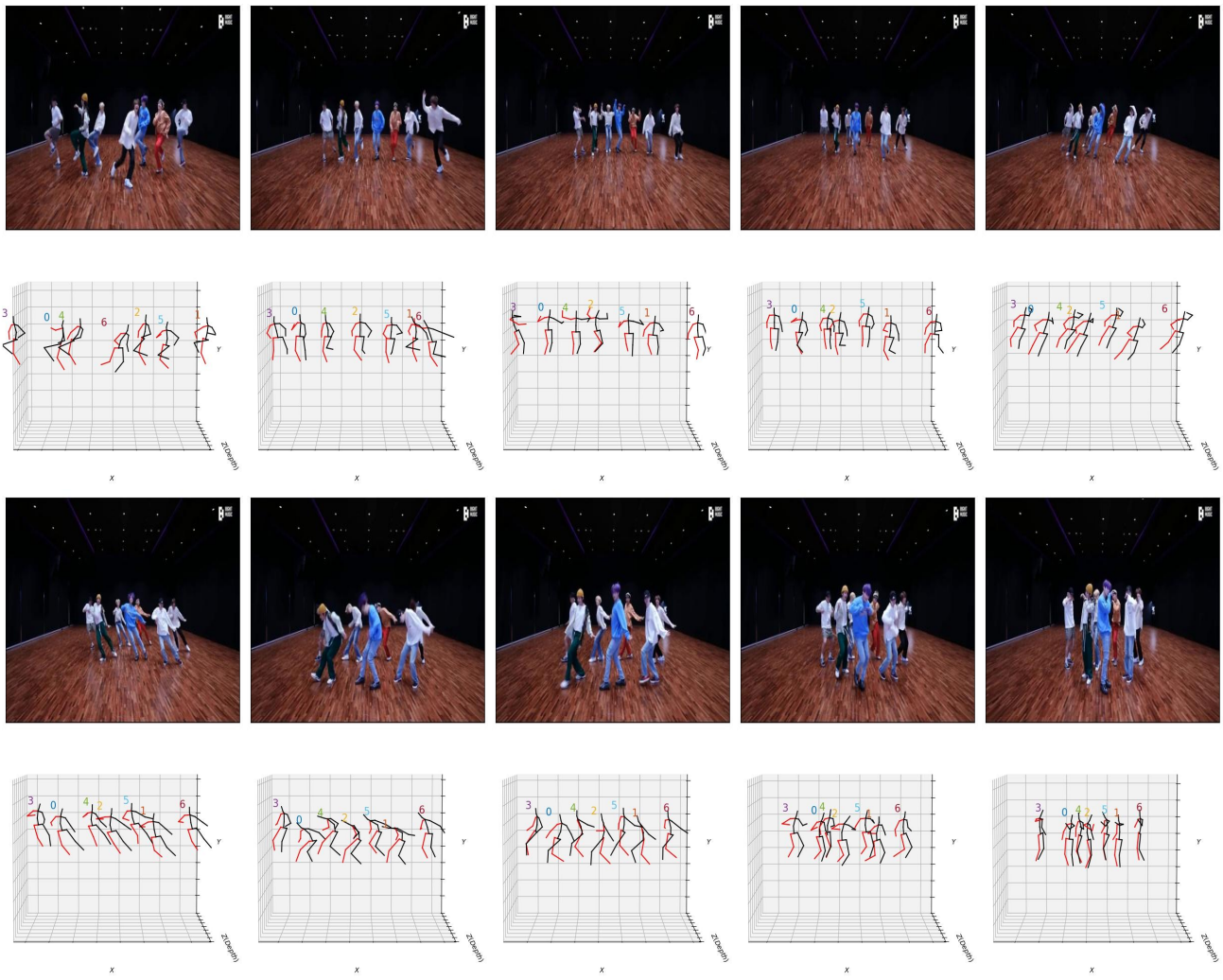

Figure IV. Additional Examples of in-the-wild inference (1/8) – Group dance **(Massive movements and occlusions)**

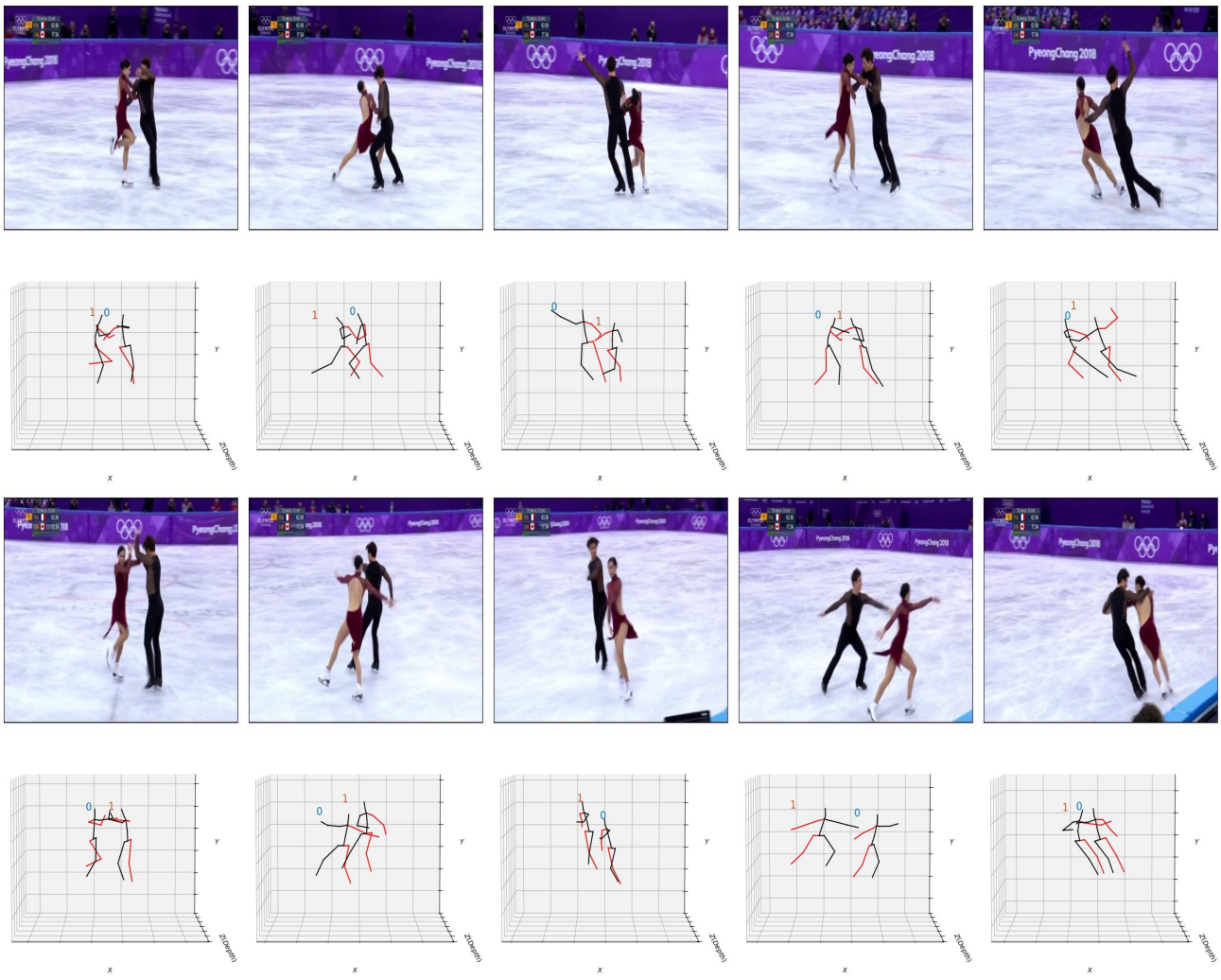

Figure V. Additional Examples of in-the-wild inference (2/8) – Figure skating **(Rampant movements and occlusions)**

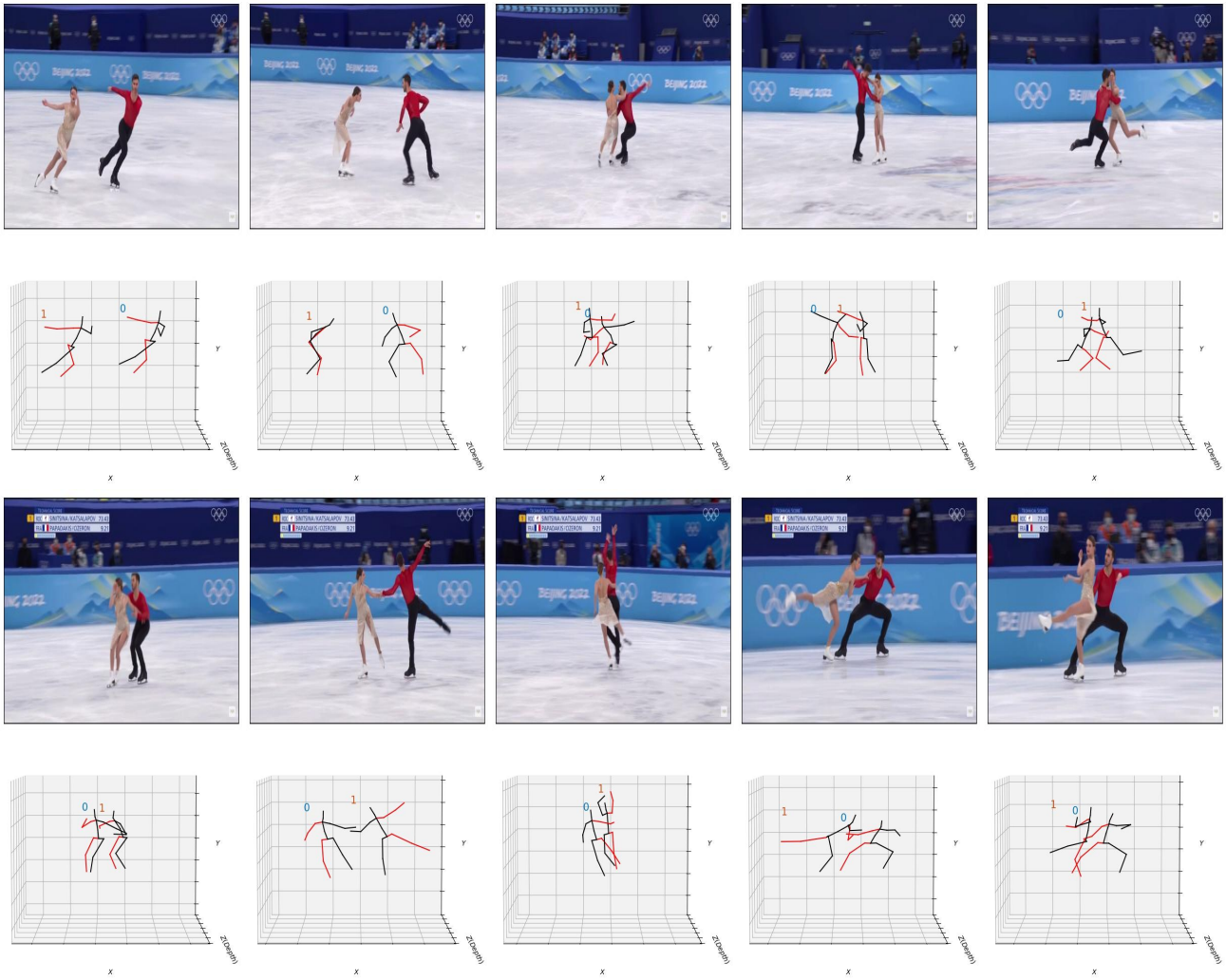

Figure VI. Additional Examples of in-the-wild inference (3/8) – Figure skating **(Rampant movements and occlusions)**

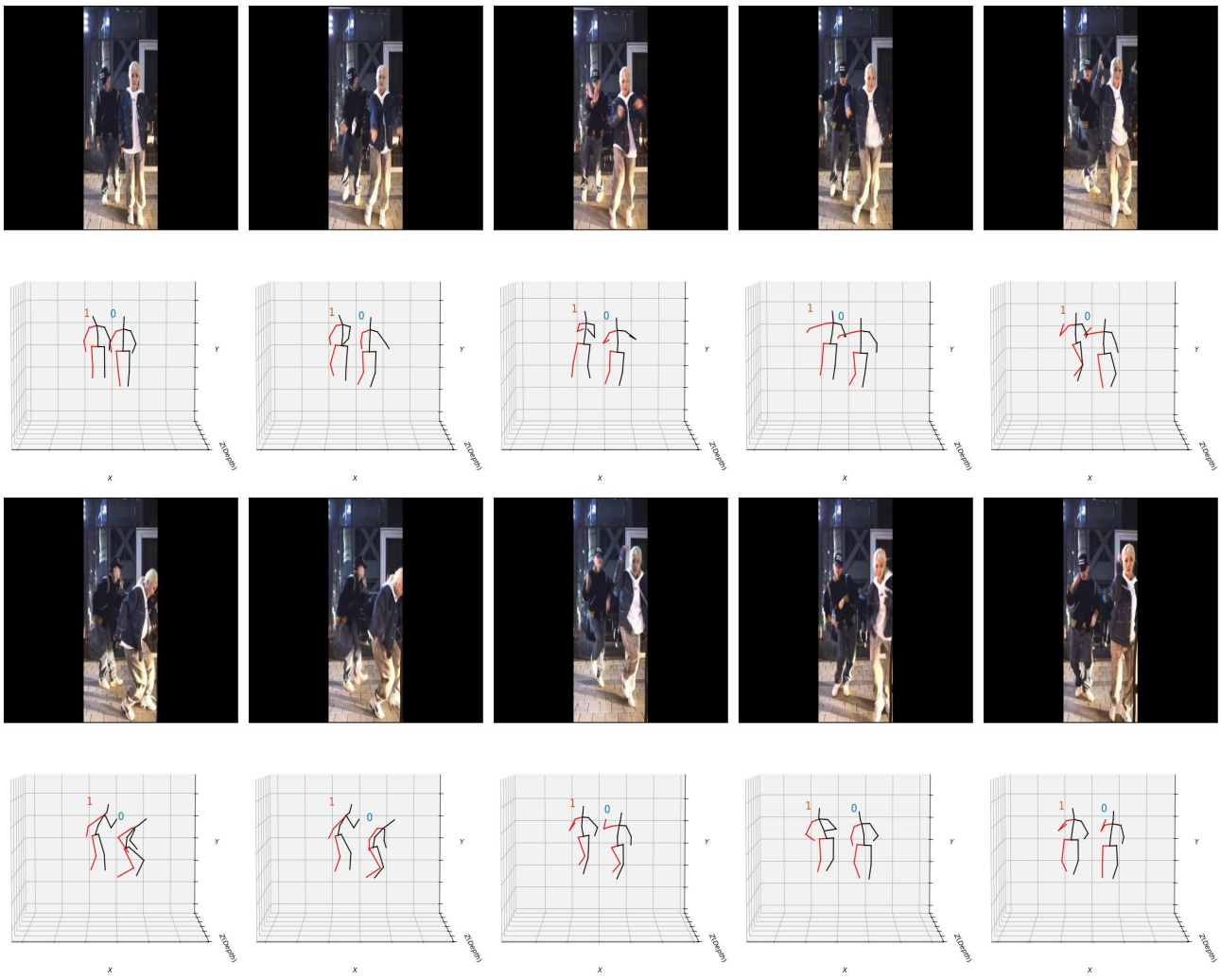

Figure VII. Additional Examples of in-the-wild inference (4/8) – Dance practicing **(Padded input)**

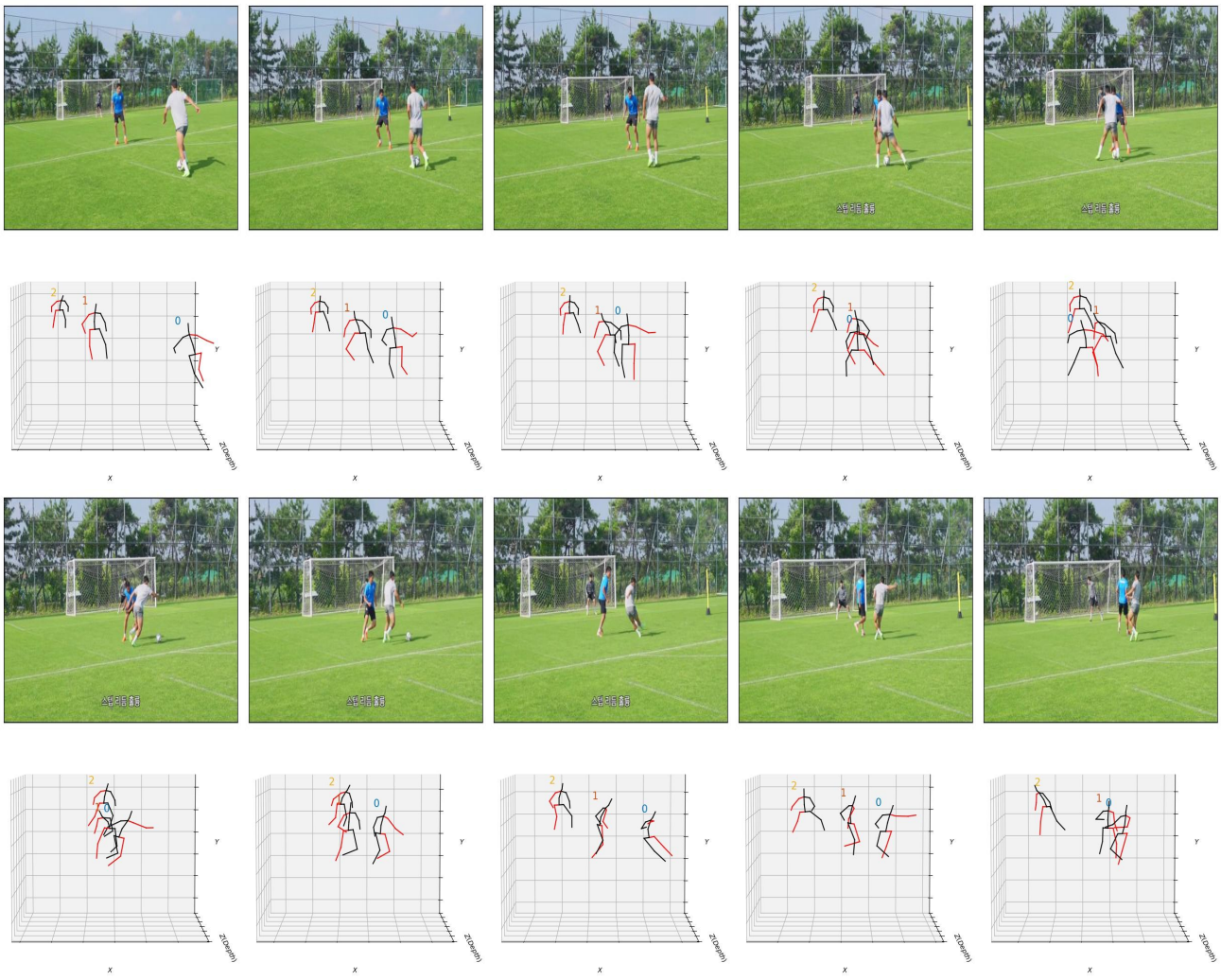

Figure VIII. Additional Examples of in-the-wild inference (5/8) – Professional soccer **(Rapid camera view change)**

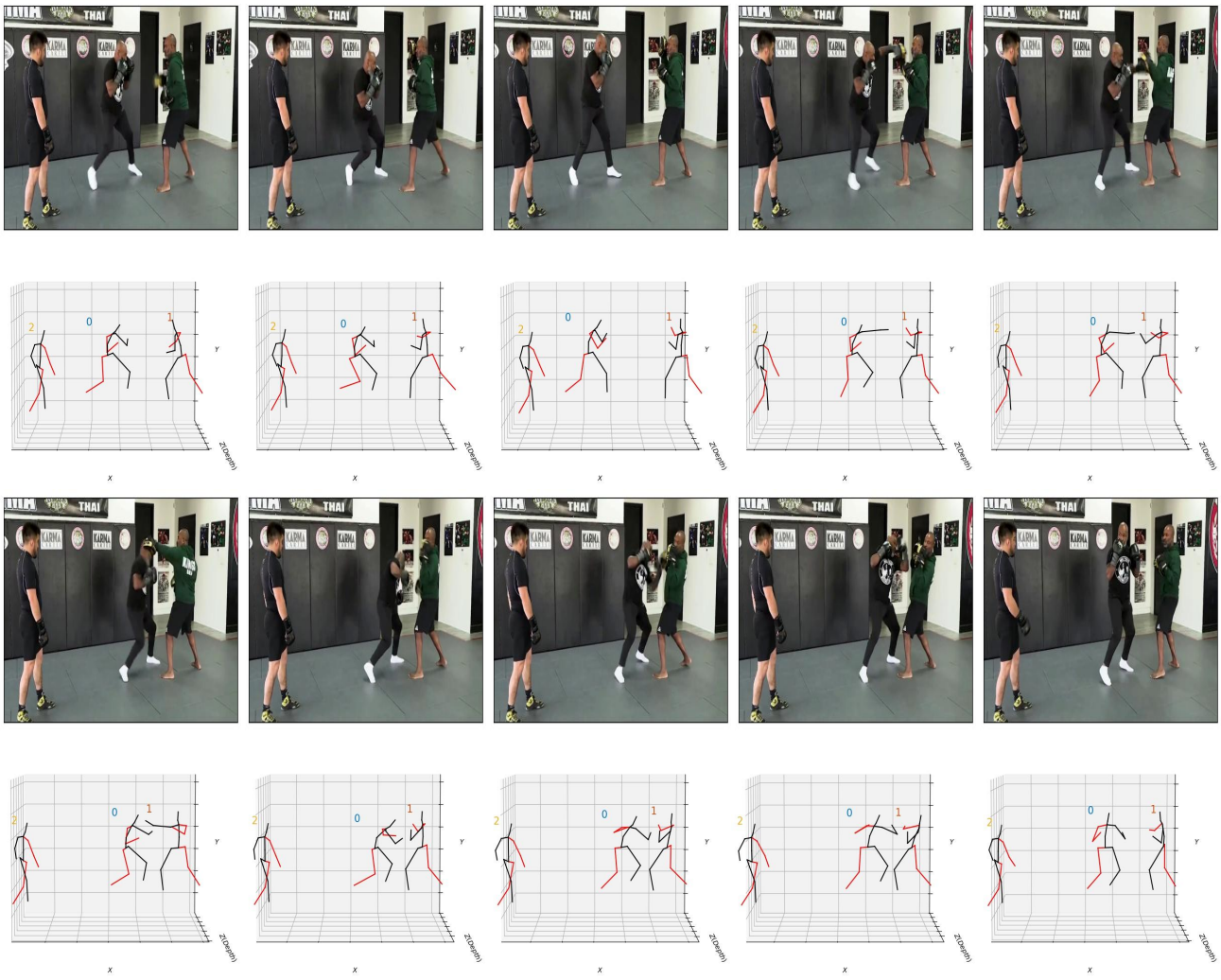

Figure IX. Additional Examples of in-the-wild inference (6/8) – Professional Boxing **(Rampant movements)**

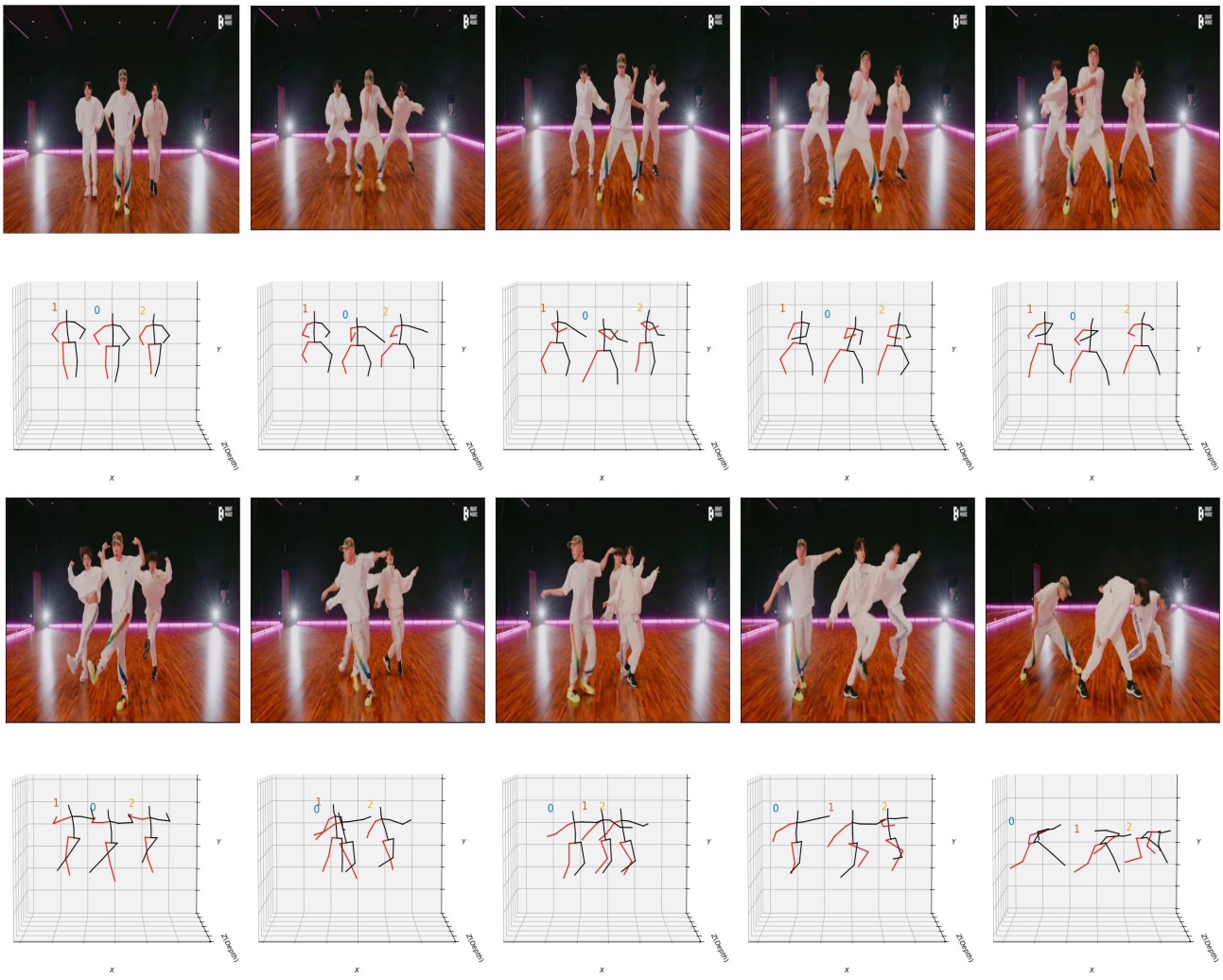

Figure X. Additional Examples of in-the-wild inference (7/8) – Boy group dance practicing **(Homogeneous Looking)**

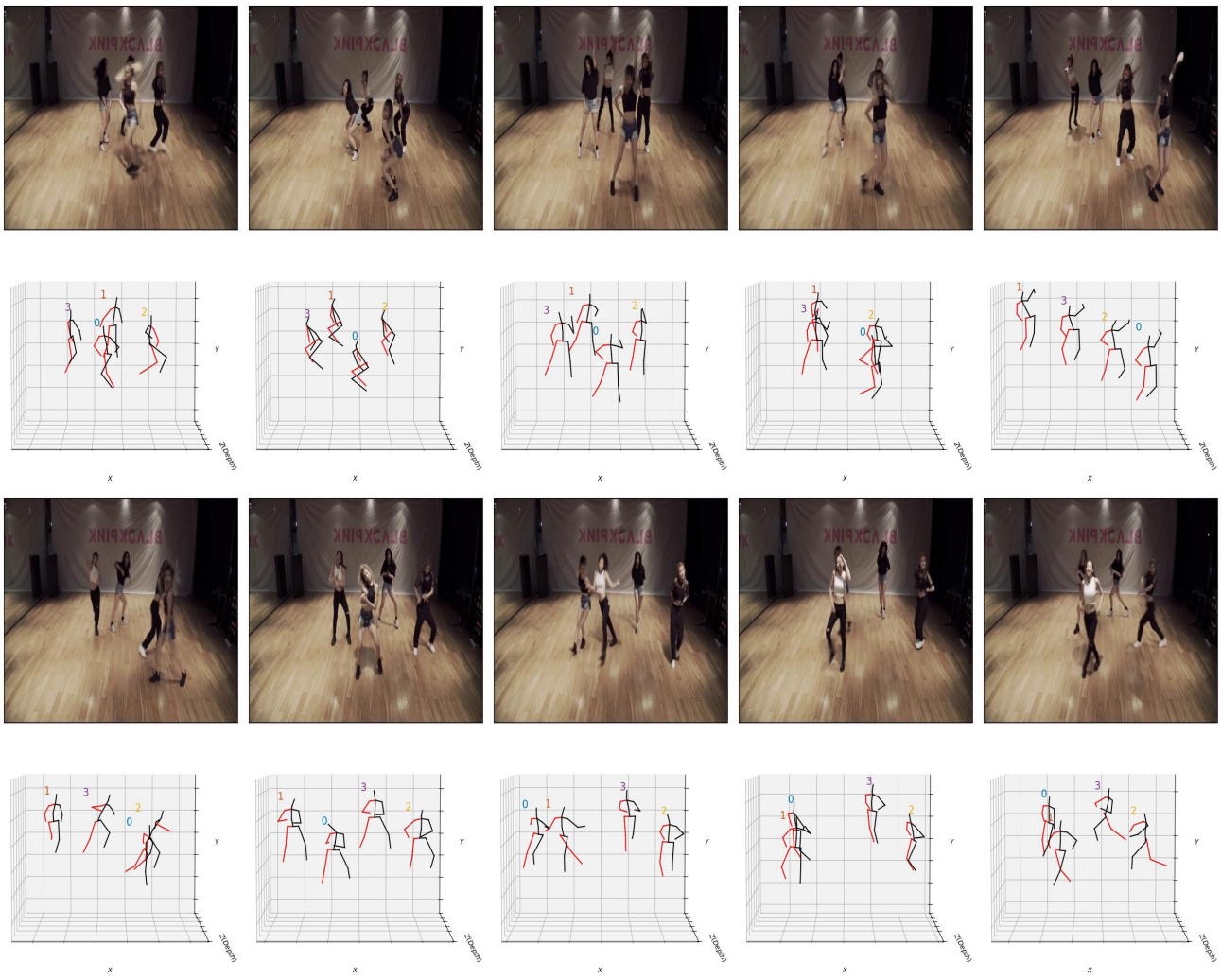

Figure XI. Additional Examples of in-the-wild inference (8/8) – Girl group dance practicing **(Massive movements and occlusions)**

Figure XII. **Failure Cases**. **(Top)** Tracking Failure, where one person is not totally tracked due to heavy occlusion. **(Bottom)** Depth Ambiguity, where the depth for children is wrongly estimated which can be checked in the side view.