1. Comparison to Other Human Datasets

Table 1 provides a comparative analysis of various human datasets, categorized as real or synthetic and captured from either ground or aerial perspectives. Key observations from this comparison are outlined below:

1. Aerial-view sets, *thanks to their wide viewing angles*, generally have *more human instances per image* than ground-view sets, except for few cases that employed a fixed number of actors in a real set or designing one instance per image in a synthetic set.
2. Aerial-view sets generally contain a wider range of viewpoints. (mostly near∼far)
3. For existing synthetic datasets, aerial-view sets typically feature *fewer motion variations* compared to ground-view sets. This is because *aerial-view datasets often prioritize leveraging a wide range of viewpoints over expanding the variety of human motions*.
4. Rule-guided design, which is only leveraged in the SynPlay, can utilize *significantly larger range of human motions* compared to detail-guided design.

The comparison shown in the table also demonstrates that SynPlay successfully addresses the shortfall of aerial-view synthetic datasets (3rd observation), while maximizing the benefits of aerial-view datasets (1st and 2nd observations).

Moreover, the 4th observation supports that our proposed rule-guided design is successful in securing the diversity of human motions in the set. It is noteworthy that while SURREAL [23] (constructed with 'detail+mocap') contains a comparable number (6.5M) of human instances as SynPlay, the number of motions manifested in the dataset is extremely limited when compared to SynPlay (23 vs. uncountable).

2. SynPlay Statistics

Here, we provide several statistics from the SynPlay. Fig 2 shows the distribution of bounding box sizes over human instances captured by each device. The majority of bounding box sizes are small, which illustrates a characteristic of aerial-view datasets. Interestingly, UAVs can capture human instances with larger bounding boxes than CCTVs. This could be due to the fact that, although UAVs are typically positioned at higher altitudes than CCTVs, there are more cases where UAVs get closer to real-time events and human instances, different from the fixed CCTVs.

Fig 3 shows the distribution of human height with respect to gender and age. As mentioned in the main manuscript, each distribution is formed as being bell-shaped. We create 456 virtual characters by controlling human height, gender, and age according to these distributions and uniquely diversifying other factors (skin color, obesity, hair, outfit, *etc*) as much as possible, as shown in Fig 1.

3. Implementation Details

3.1. Motion evolution graph

Fig 4 shows motion evolution graphs used in designing the game scenarios for the SynPlay dataset. Even within the same game, the scenario may change, but the motion evolution graph will remain consistent. It is noteworthy to mention that, despite the wide range of situations and a variety of motions involved in the games, the motion evolution graph for each game consists of only a few motion nodes and their transitions. Given that each node (represented as an *elementary motion state* in the main manuscript) encompasses a range of motions, this illustrates the essence of a rule-based design approach where only basic game rules are provided to freely allow the diverse array of human motions to be manifested.

3.2. Experimental setting

In our experiments, our goal is to explore the capabilities of SynPlay as supplementary training data on a variety of tasks. We mostly adhere to the original settings and implementations of the methods used in our experiments, with minimal modifications. The specific modifications used in our experiments are described below.

Architecture modification. Our tasks, specifically human detection and semantic segmentation, can be viewed as one-class problems. Therefore, all method architectures, particularly the dimensions of the last layer, have been adjusted accordingly.

Image size applied in YOLOv8 training/inference. We use the image size of 1280×1280 for all datasets except for COCO, which uses an image size of 640×640. This decision simply takes into account the original image size of the datasets. Even after rescaling, the size range of human instances in the compared datasets remains similar. When using other models, *i.e.,* Mask2Former in semantic segmentation tasks and retinaNet in data-scarce tasks, the image size/scaling recommended in the original settings was used.

Training Mask2Former without the large-scale jittering (LSJ) augmentation [9]. We did not use the default LSJ augmentation in training the Mask2Former segmentation models solely for performance reasons. In all cases, segmentation accuracy were found to be significantly lower when LSJ augmentation was used. LSJ augmentation, which greatly expands the range of image scaling, may not be suitable for aerial-view detection, which mainly includes small-sized human instances. This performance degradation with LSJ augmentation is also observed in [10], which is a reputable literature in the field of self-supervised learning.

Settings for PT-FT. When using PT-FT in the general tasks, training specifications, including training epochs and learning rate, did not differ between pre-training and fine-tuning.

Table 1. **Comparison of human datasets.** '#inst/img' is acquired only on images that contain human. '#motion' indicates the number of unique motions depicted in the dataset, except the ones with the subscript 'pose' which indicate the number of static poses. Since a single motion can consist of multiple number of unique poses, #motion is generally smaller than the number of poses. For certain datasets, the test set without available labels is excluded from this comparison. 'uncountable' indicates that the number of human motions included in the set is countless/uncountable.

dataset	domain	#inst	#img	#inst/img	natural motion	#motion	viewpoint
ground-view							
VOC `12` [7]	real	10K	11.5K	2.48	daily	uncountable	near
KITTI [8]	real	4.5K	7.5K	2.52	daily	2	near
COCO `Dev17` [13]	real	649K	164K	9.72	daily	uncountable	near
MPII Human Pose [2]	real	40K	24.9K	1.61	daily	20	near
Cityscapes [6]	real	21.4K	5K	7.85	daily	2	near
ADE20K [26]	real	30K	27.5K	4.36	daily	uncountable	near
Human-Art [12]	real	123K	50K	2.46	art	uncountable	near
GTA5 [16]	synth	1.4M	1.4M	1	✗	$20K_{\text{pose}}$	near
SURREAL [23]	synth	6.5M	6.5M	1	detail+mocap	23	near
SOMAset [3]	synth	100K	100K	1	detail+mocap	250_{pose}	near
PersonX [22]	synth	273K	273K	1	✗	4_{pose}	near
UnrealPerson [25]	synth	120K	120K	1	✗	2	near
AGORA [15]	synth	·	19K	1∼15	detail+mocap	$4,240_{\text{pose}}$	near
BEDLAM [5]	synth	·	380K	1∼10	detail+mocap	$2,311_{\text{pose}}$	near
aerial-view							
Okutama-action [4]	real	·	77K	∼9	detail	12	med
Semantic Drone [1]	real	1.5K	400	4.16	daily	unspecified	med
UAVid [14]	real	4.7K	420	20.06	daily	unspecified	med∼far
VisDrone [27]	real	109K	40.0K	15.42	daily	unspecified	med
Archangel-real [20]	real	165.6K	41.4K	4	detail	3_{pose}	near∼far
Archangel-mannequin [20]	real	·	178.8K	6∼7	detail	3_{pose}	near∼far
Archangel-synth [20]	synth	4.4M	4.4M	1	✗	3_{pose}	near∼far
SynDrone [17]	synth	803K	72K	11.15	✗	2	med∼far
CARGO [24]	synth	108K	108K	1	✗	2	near∼far
SynPlay	synth	6.5M	73K	88.40	rule+mocap	uncountable	near∼far

* natural motion
· daily: human motions engaged in daily activity
· art: human motions shown in works of art
· detail: human motions captured by 'detail-guided design'
· rule: human motions captured by 'rule-guided design'
· +mocap: human motions captured using a motion scanner

In data-scarce tasks, we follow all the settings of [19] as outlined in PTL, while leaving out the progressive component.

Settings for data-scarce tasks. For all experiments performed for data-scarce tasks including the scaling behavior study, we follow all the settings and experimental environments of [21].

3.3. Quantitative measures

We provide the detail on how we calculate the two metrics used for the quantitative analysis in the main manuscript.

Fréchet Inception Distance (FID) [11]. We utilized the PyTorch implementation of FID in [18] with the default setup to assess the fidelity and diversity for all the training datasets involved in our experiments. We did not perform image scaling on the input for any dataset, and the final average pooling features were used to compute FID.

Figure 1. **456 virtual players in SynPlay** created using Character Creator.

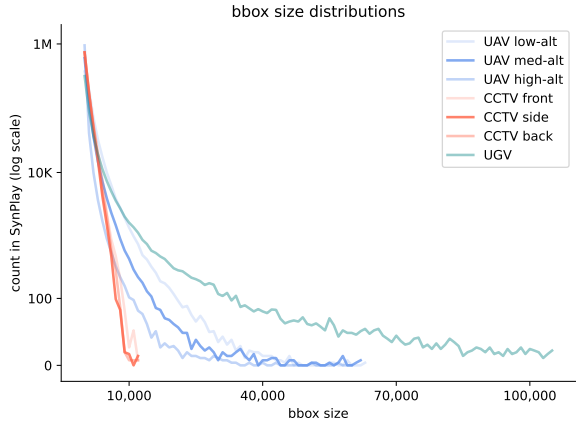


Figure 2. **Bounding box size distribution** for each image-capturing device.

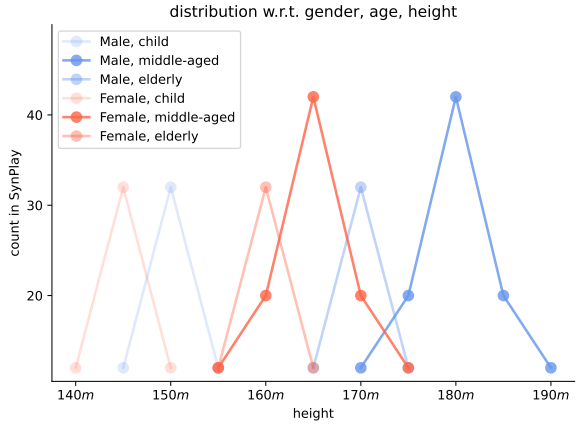


Figure 3. **Character height distribution** that varies according to gender and age.

Proportion of nadir-view instances. An instance with an elevation angle greater than 71.57° relative to the UAV is considered to be a nadir-view instance, representing the maximum elevation angle for Archangel* [21]. To identify nadir-view instances for Archangel, we utilized the dataset metadata, i.e., UAV position. Similarly, for Archangel*, we determined if an instance was an nadir-view instance based on the source instance, also using the dataset metadata. In the case of SynPlay, we computed the elevation angle for each instance using the absolute 3D coordinates of the instance and the UAV provided by SynPlay.

4. Qualitative Analysis

4.1. Blending and animation layer

Fig 5 and 6 show several examples of the blending process and how the animation layers are leveraged: the two techniques for expanding human motions within the virtual environments, respectively. Interestingly, the motions created by blending is largely different from their corresponding input motions, while the motions created via leveraging the animation layers still exhibit the appearances and dynamics resembling both the input motions. These two techniques are readily available for use within the Unity environment.

4.2. Virtual motion and real-world motion

Fig 7 shows several examples of real-world motions. Real-world motions are created either by having the real human wearing the motion capture device mimic the pre-provided reference motions or by demonstrating potential in-game motions under the given game rules. It is observed that real-world motions can express a wider range of specific actions while maintaining a sense of realism. Moreover, motions that are difficult to pinpoint or describe can also be created,

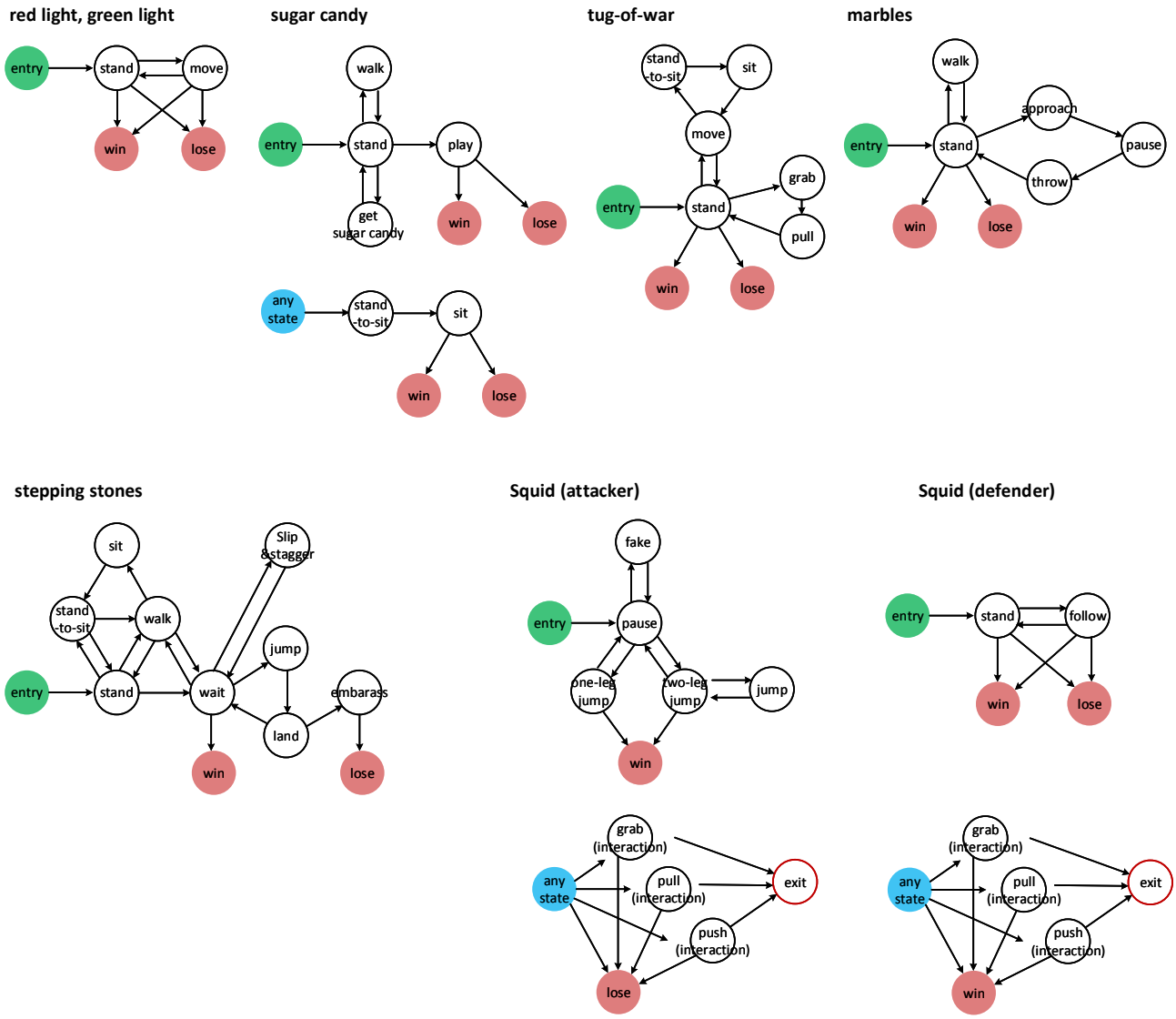


Figure 4. **Motion evolution graphs.** The start node ('entry') and the end nodes ('win' or 'lose') are indicated by green and red circles, respectively. For the games where secondary graphs are available (i.e., sugar candy or squid), at any given time (except at the start or end node), the current state in the main graph can move to the 'any state' node (blue-filled circle) in the secondary graph. When the 'end' node (red-bordered circle) is reached within the secondary graph, the current state moves its way back to the latest node that was touched in the main graph before entering the secondary graph.

e.g., multi-person wrestling motions.

4.3. SynPlay sample images

Fig 8 includes additional sample images from the SynPlay dataset. Various human appearances depending on human motion differently taken according to the game scenario, and camera viewpoints are observed. Various human appearances are observed that change depending on human motions taken differently according to the game scenario, and different camera viewpoints. In addition, various characters and backgrounds used for creating SynPlay are also visible.

5. Benchmark Aerial View Dataset

Fig.9 presents qualitative detection results acquired on representative aerial-view human datasets, highlighting the fundamental differences between aerial and ground-view perception. Unlike conventional datasets such as MS COCO [13], aerial human imagery involves small-scale human instances, often spanning only tens of pixels, and is captured from extreme and diverse viewpoints, including

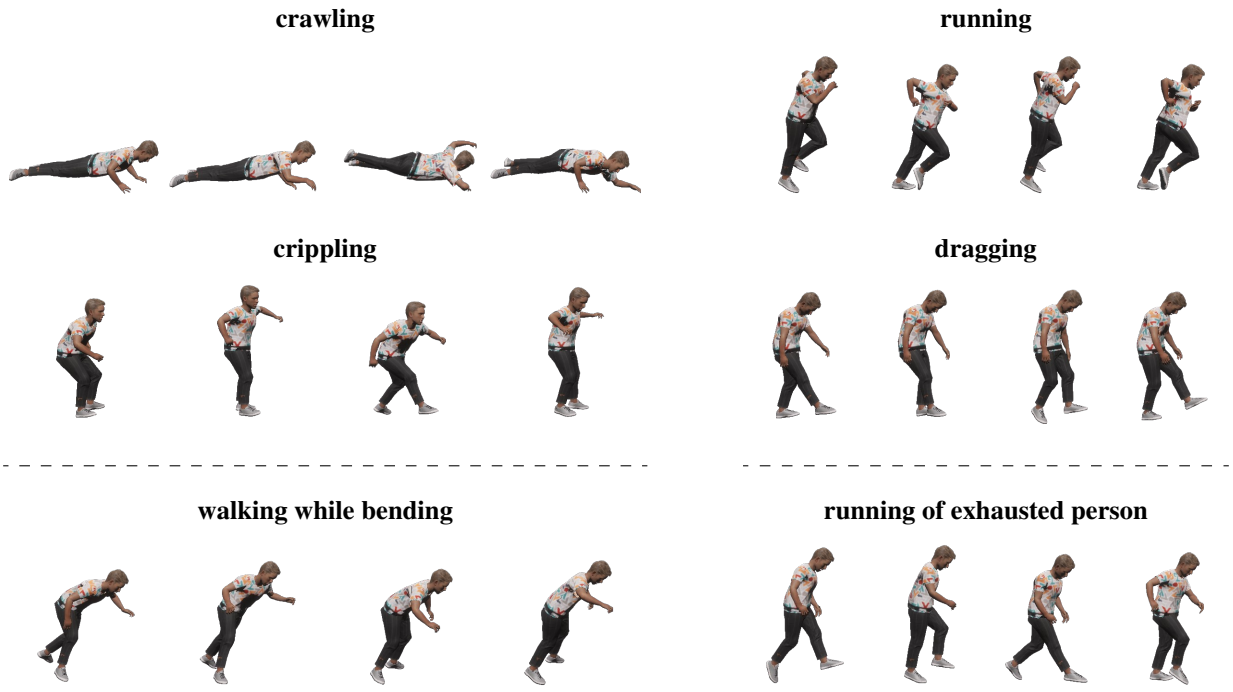


Figure 5. **Two motion blending examples**. For each example (left or right column), the two motions (top and middle row) is blended together to generate a new motion (bottom row). The blending ratio between the two input motions can be controlled. The names for the motions are not computationally involved in the blending process.

nadir, oblique, and off-nadir perspectives.

These factors create distinct challenges: while detailed appearance cues like facial features or textures become less informative at such scales, *motion patterns, postures, and interaction dynamics* remain critical. This requires datasets that concurrently capture both motion diversity and viewpoint variation.

SynPlay directly addresses these challenges by combining *multi-perspective captures* with *rule-guided behavioral diversity*, enabling robust human identification in aerial scenarios where existing datasets fall short as shown in the qualitative comparisons.

6. Limitations and Future Directions

A significant number of human instances in SynPlay appear at very low resolutions, which is an inherent challenge in any aerial-view dataset. However, as shown in the bounding box size histogram included in the supplementary material, SynPlay also contains many high-resolution human instances, with approximately 10,000 examples having bounding box areas greater than 10,000 pixels. We encourage selectively using these instances depending on the requirements of specific tasks.

SynPlay was developed to provide rich visual representations of human appearance for tasks focused on localizing people in complex scenes. While it centers on the human domain, we recognize the value of incorporating features from a broader set of object categories. Expanding SynPlay to include a wider array of objects could further enhance its utility for training and evaluating general-purpose models.

Future directions for the community include improving photorealism, simulating sensor artifacts such as rolling shutter and motion blur, and integrating synthetic and real data through hybrid training protocols. Adjusting scenario priors to support more diverse interactions and cultural settings, expanding metadata for occlusion and weather conditions, and incorporating additional sensing modalities like thermal imagery may also support greater robustness. More detailed behavioral annotations could further enable research in tracking, forecasting, and social interaction understanding. These extensions build on SynPlay's strengths in controllability, scale, and behavioral diversity while helping shape the next generation of aerial perception benchmarks.

References

[1] Aerial semantic segmentation drone dataset. http://dronedataset.icg.tugraz.at. 2, 12

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. CVPR*, 2014. 2

[3] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, and Theoharis Theoharis. Looking be-

Figure 6. **Three examples of leveraging animation layers**. For each example (left, middle, or right), the resulting motion of leveraging the animation layers over two input motions (top and middle rows) is shown in the bottom row. Note that the semantic labels (e.g., walking, cheering) were not provided at the time of capture; they are included in the figure only for the convenience of the readers.

yond appearances: Synthetic training data for deep CNNs in re-identification. *Comput. Vis. Image Underst.*, 167:50–62, 2018. 2

[4] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proc. CVPR Workshop*, 2017. 2, 12

[5] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proc. CVPR*, 2023. 2

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson ad Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. CVPR*, 2016. 2

[7] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015. 2

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proc. CVPR*, 2012. 2

[9] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proc. CVPR*, 2021. 1

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, 2022. 1

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. NeurIPS*, 2017. 2

[12] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-Art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proc. CVPR*, 2023. 2

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV*, 2014. 2, 4, 12

[14] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. UAVid: A semantic segmentation

dataset for uav imagery. *ISPRS J. Photogramm Remote Sens. (P&RS)*, 165:108–119, 2020. 2

[15] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proc. CVPR*, 2021. 2

[16] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proc. ECCV*, 2016. 2

[17] Giulia Rizzoli, Francesco Barbato, Matteo Caligiuri, and Pietro Zanuttigh. Syndrone-multi-modal UAV dataset for urban scenarios. In *Proc. ICCV Workshop*, 2023. 2

[18] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 2

[19] Yi-Ting Shen, Hyungtae Lee, Heesung Kwon, and Shuvra S. Bhattacharrya. Progressive transformation learning for leveraging virtual images in training. In *Proc. CVPR*, 2023. 2

[20] Yi-Ting Shen, Yaesop Lee, Heesung Kwon, Damon M. Conover, Shuvra S. Bhattacharyya, Nikolas Vale, Joshua D. Gray, G. Jeremy Leongs, Kenneth Evensen, and Frank Skirlo. Archangel: A hybrid UAV-based human detection benchmark with position and pose metadata. *IEEE Access*, 11:80958–80972, 2023. 2

[21] Yi-Ting Shen, Hyungtae Lee, Heesung Kwon, and Shuvra S. Bhattacharyya. Diversifying human pose in synthetic data for aerial-view human detection. arXiv:2405.15939, 2024. 2, 3

[22] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proc. CVPR*, 2019. 2

[23] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proc. CVPR*, 2017. 1, 2

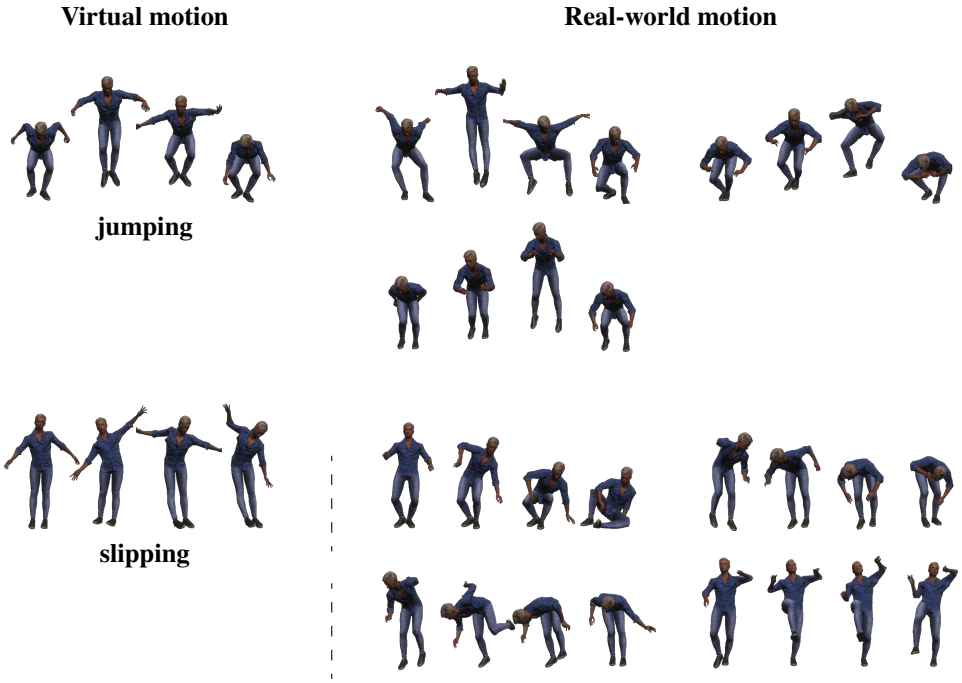
[24] Quan Zhang, Lei Wang, Vishal M. Patel, Xiaohua Xie, and Jianhuang Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *Proc. CVPR*, 2024. 2

[25] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. UnrealPerson: An adaptive pipeline towards costless person re-identification. In *Proc. CVPR*, 2021. 2

[26] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proc. CVPR*, 2017. 2

[27] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7380–7399, 2022. 2, 12

(a) Using virtual motions as reference



(b) Without reference

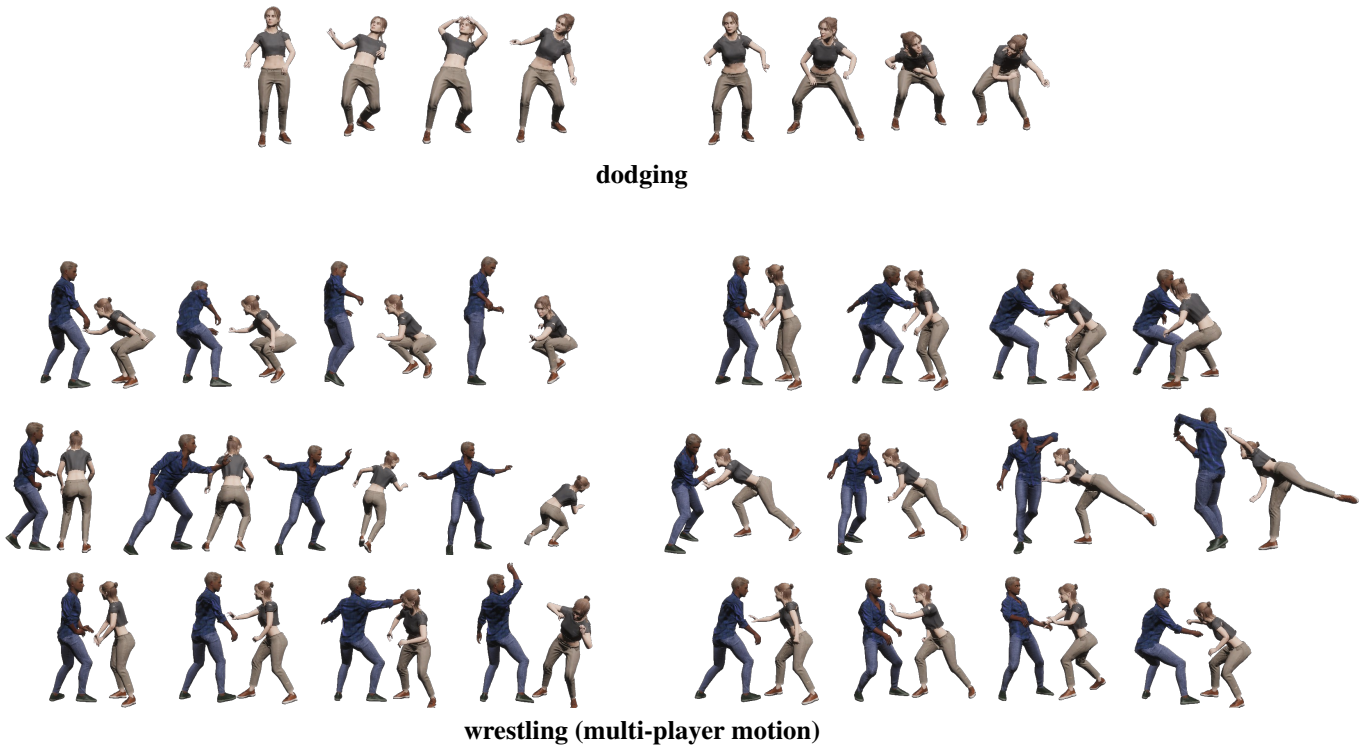


Figure 7. **Real-world motion examples.** Real-world motions are acquired either (a) by mimicking reference motions or (b) by exhibiting potential in-game motions without any reference that align with the given game rules. Wearable motion scanners are used for all the cases. Note that the semantic labels (e.g., jumping, dodging) were not provided at the time of capture; they are included in the figure only for the convenience of the readers.

(1) red light, green light

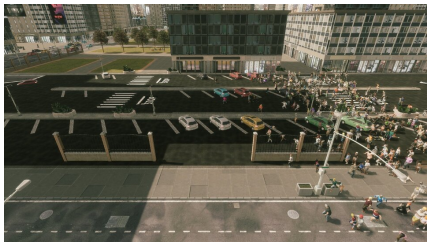
UGV

UAV med-alt

UAV low-alt

CCTV back

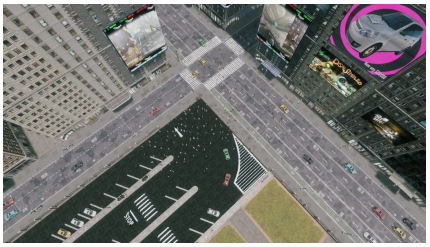
UAV high-alt



CCTV side

(2) sugar candy

CCTV side



UAV high-alt

UGV

UAV low-alt

CCTV front

CCTV back

(3) tug-of-war

UAV low-alt

CCTV front

UAV high-alt

CCTV back

UAV low-alt

UGV

(4) marbles

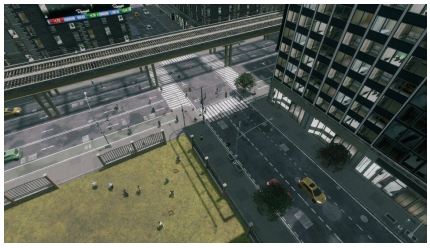
UGV

CCTV front

UAV med-alt

CCTV side

UAV med-alt



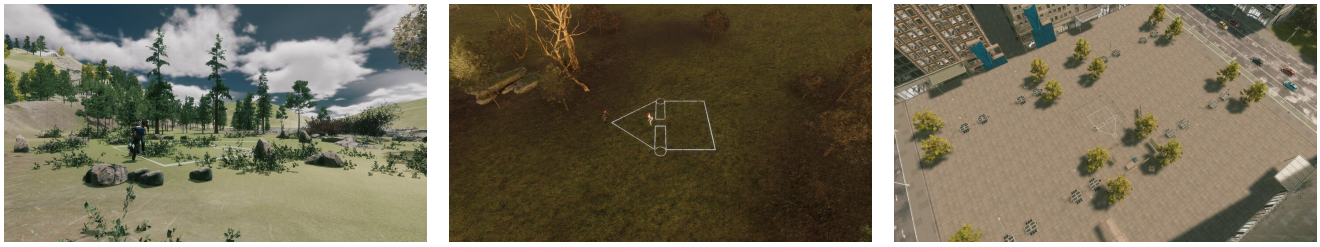
UAV low-alt

(5) stepping stones

CCTV front | UAV med-alt | UAV high-alt

UGV | UAV med-alt | CCTV side

(6) squid



UGV | CCTV front | UAV high-alt

CCTV side | UGV | UAV low-alt

Figure 8. **More example images from SynPlay** are shown for all six Korean traditional games, each with various camera viewpoints.

VisDrone [27] Okutama-action [4] SemanticDrone [1]

Figure 9. **Qualitative detection samples for baselines vs. SynPlay+real models.** Baseline models are trained on **real** dataset. □ (green bounding boxes) indicate baseline (**real only**) human detection outputs while □ (magenta bounding boxes) indicate the ones acquired by the models trained on corresponding **SynPlay+real** dataset. Note that unlike conventional ground-level benchmarks such as MS COCO [13], aerial-view perception involves **small-scale human instances**, **extreme viewpoint variations**, and unique appearance challenges due to camera altitude and perspective. This highlights the need for specialized synthetic datasets like SynPlay that are tailored for long-range aerial human analysis.