On the Equivalency, Substitutability, and Flexibility of Synthetic Data

Che-Jui Chang
Rutgers University
chejui.chang@rutgers.edu

Danrui Li
Rutgers University
danrui.li@rutgers.edu

Seonghyeon Moon
Rutgers University
sm206@cs.rutgers.edu

Mubbasir Kapadia
Roblox
mkapadia@roblox.com

Abstract

We study, from an empirical standpoint, the efficacy of synthetic data in real-world scenarios. Leveraging synthetic data for training perception models has become a key strategy embraced by the community due to its efficiency, scalability, perfect annotations, and low costs. Despite proven advantages, few studies put their stress on how to efficiently generate synthetic datasets to solve real-world problems and to what extent synthetic data can reduce the effort for real-world data collection. To answer the questions, we systematically investigate several interesting properties of synthetic data – the equivalency of synthetic data to real-world data, the substitutability of synthetic data for real data, and the flexibility of synthetic data generators to close up domain gaps. Leveraging the M^3`Act` *synthetic data generator, we conduct experiments on DanceTrack and MOT17. Our results suggest that synthetic data not only enhances model performance but also demonstrates substitutability for real data, with 60% to 80% replacement without performance loss. In addition, our study of the impact of synthetic data distributions on downstream performance reveals the importance of flexible data generators in narrowing domain gaps for improved model adaptability.*

1. Introduction

For the past decade, collecting real-world data with human annotations has been driving the momentum of research advancements in the field of computer vision and machine perception. Despite the tremendous breakthrough in many tasks, obtaining these datasets with high-quality human annotations remains costly and labor-intensive, thus hindering the development for certain research tasks that require fine-grained annotations, including human pose and shape estimation[3, 15], multi-object tracking [16, 22, 27, 28], and collective activity understanding [8–11, 13, 17, 20, 29].

The adoption of synthetic data from game engines has become an emerging alternative to real-world data, due to its efficiency, flexibility, scalability, and perfect annotations. Previous studies [3, 12, 15, 23, 26] also show the capability of synthetic data in providing diverse and photorealistic images, generating customized datasets for edge cases, improving benchmark performances on downstream datasets, and mitigating ethical risks, for a wide range of vision and perception tasks. Nonetheless, related studies are rarely focused on the similarity between data distributions, the equivalency of synthetic data to real-world data, the substitutability of synthetic data for real data, and the flexibility of synthetic data generators to close up the domain gap. Therefore, it remains unclear to the community how to efficiently generate synthetic datasets to solve real-world problems and to what extent synthetic data can benefit real-world tasks.

In this work, we aim to study the aforementioned properties of synthetic data, by systematically experimenting with different data sources for training and comparing the performance of models trained on synthetic data to those trained on real-world data. We adopt the multi-person tracking (MPT) task as the focus of our study and leverage M^3`Act` [12] as the synthetic data generator. M^3`Act` features multiple semantic groups and produces highly diverse and photorealistic videos with a rich set of annotations suitable for human-centered tasks, including multi-person tracking, group activity recognition, and controllable human group activity generation. Notably, it offers a flexible data generation pipeline to modify the data distribution by editing the highly parameterized and modularized human groups, enabling us to examine the influence of synthetic data distributions on downstream task performances.

The key contribution of this work lies in the proven efficacies and insights gained from our investigations of the equivalency of synthetic data to real-world data and the substitutability of synthetic data for real data. We show that synthetic data can not only improve model performance on downstream datasets but also effectively replace

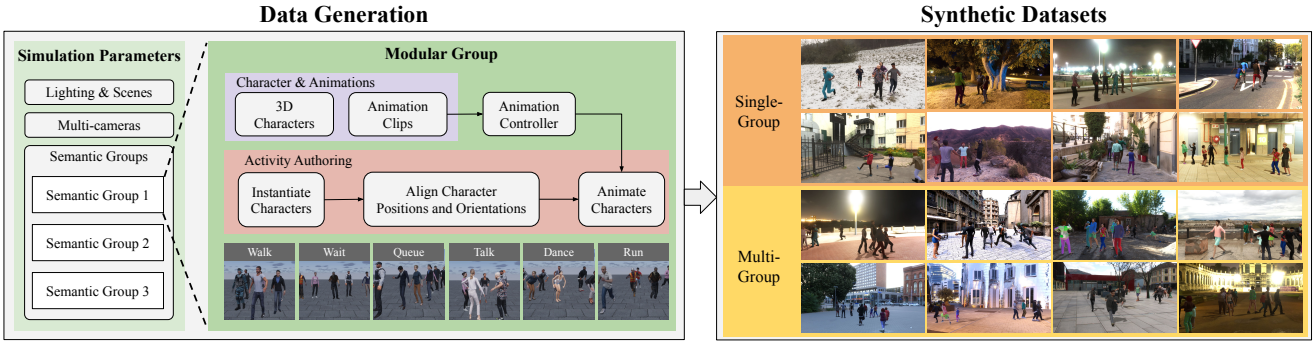

Figure 1. The data generation process of M^{3}Act. The data generation process is highly parameterized, enabling the adjustment of resulting synthetic data distributions, as illustrated at the right.

up to 80% of MOT17 data [19] without compromising performances. Moreover, our investigation of the impact of synthetic data distributions on downstream model performances highlights the importance of a flexible synthetic data generator to shorten domain gaps, thereby enhancing the adaptability of models trained on synthetic data to real-world scenarios

2. Background

Synthetic data plays a vital role in contemporary computer vision and machine learning research, particularly because of the increasing demand driven by large data-hungry models [5, 21]. The community widely acknowledges the benefits of employing synthetic data for training machine learning models, due to its efficiency, enhanced performance, customizability, and cost reduction. A key characteristic of synthetic data is its task-specific nature, where datasets are tailored to specific research objectives or scenarios. This customization allows researchers to design data generators to produce synthetic data that closely aligns with the requirements of the downstream experiments, enhancing the performance of machine learning models trained on synthetic data. For example, SURREAL [23] was designed for human pose estimation to improve the background and data diversity. BEDLAM [3] integrates physically animated clothes into its data generation pipeline for improved performance on human pose and shape estimation. M^{3}Act [12] leverages the rule-based authoring of human group activities for multi-person and multi-group research. The same study also shows that datasets such as BEDLAM [3] and GTA-Humans [6] are inadequate for multi-person tracking.

Despite these advantages, prior studies have primarily focused on the enhanced performances achieved with synthetic data. While performance enhancement is crucial, seeking optimality may not always be practical in real-world scenarios. Understanding the tradeoffs involved in collecting more data and the extent to which it leads to performance improvements is equally important. Previous studies [12] have hinted at the potential for synthetic data to replace certain amounts of real data, yet these properties of synthetic data have not been thoroughly explored. In this study, we delve into several interesting properties – the equivalency, substitutability, and flexibility of synthetic data. We demonstrate through experiments how these properties can provide valuable insights for using synthetic data in real-world scenarios.

3. Overview of M^{3}Act

We provide an overview of the data generation process, the flexible parameterization, and the properties of M^{3}Act.

3.1. Synthetic Data Generation

M^{3}Act is a synthetic data generator built with Unity Engine [4]. It features multiple semantic groups and produces highly diverse and photorealistic videos with a rich set of annotations suitable for human-centered tasks. The data generator contains 25 scenes, 104 HDRIs, 5 lighting volumes, 2200 human models, 384 animations, and 6 group activities. M^{3}Act is tailored to support multi-person and multi-group research, which effectively enhances the model performances on multi-person tracking and group activity recognition, and enables novel research, controllable group activity generation. Its data generation process, illustrated in Fig. 1, is procedurally driven and contains a high degree of simulation parameters, including scene parameters like lighting, scenes, and cameras, as well as group parameters such as animations, characters, alignments, and group types. This level of flexibility and customization is rarely seen in previous synthetic data works [1, 3, 23, 26]. It allows for the adjustment of the simulation parameters within the generator, enabling the customization of synthetic data distributions to match specific research requirements.

3.2. Implications

Despite the promising contributions of M^{3}Act to performance enhancement, their study has revealed initial findings on several interesting properties of synthetic data.

First, the variability in synthetic data distributions resulting from the flexible parameterization of M^{3}Act presents a unique challenge in ensuring the suitability of synthetic data for downstream tasks. This challenge is particularly pronounced in multi-person scenarios, which are significantly more complex compared to tasks addressed by previous synthetic datasets [3, 6, 26]. Second, synthetic data generated by M^{3}Act exhibits properties of equivalency and substitutability, with results indicating that synthetic data can effectively replace a substantial portion of real data for multi-person tracking. Our deep investigation of these properties extends beyond the initial findings and offers valuable insights into the understanding of these synthetic data properties in real-world scenarios.

4. Experiments

We evaluate the effectiveness of synthetic data by conducting experiments on multi-person tracking (MPT), the objective of which is to predict the tracklets of all individual persons given an input video stream. The tracking performances are measured on two distinct real-world datasets: DanceTrack [22] (DT) and MOT17 [19]. The former presents a particularly challenging multi-person tracking scenario characterized by dynamic dance movements with individuals of uniform outfits. It has a total of 100 videos with over 105K frames. The latter is a widely used dataset for multi-object tracking, with objects including pedestrians, bicycles, and cars, captured in outdoor scenes. The main challenges involve crowded scenarios and interruptions in tracking due to dynamic camera movements. The dataset contains 7 long videos with a total of 11K frames.

4.1. Preliminaries

Our experiments cover the following properties of synthetic data and research questions:

- *Equivalency* and *Substitutability*. How much synthetic data is equivalent to real data? In other words, how much real data can be substituted by synthetic data, without sacrificing performance? Answers to the question would help the community understand the practical merits of using synthetic data in reducing the cost of data collection and annotation in real-world scenarios.
- *Distribution* and *Flexibility*. How does the distribution of the synthetic data affect the performance of target datasets and can we narrow the domain gaps by adjusting the distribution of synthetic data? This helps us understand the protocol of designing and selecting useful synthetic data for enhancing downstream task performance in the target domain.

To answer the first question, we conduct the experiment on MOT17 [19] dataset. We first train the model with 100% real data as a comparison baseline. Then we use a constant amount of synthetic data for pretraining and then gradually reduce the amount of real data used for finetuning. This enables us to not only observe the changes in tracking performances but also identify how much real data can be replaced by synthetic data, by comparing the performances with the baseline. We follow the protocol of M^{3}Act [12] and use the same portion of synthetic data in this experiment. For fair comparisons, we pre-trained the model on synthetic data for 5 epochs and then fine-tuned it for 20 epochs for all conditions.

Pretrain → Finetune	HOTA↑	DetA↑	AssA↑	IDF1↑	MOTA↑
N/A → 100% MOT17	56.3	51.9	62.1	66.6	56.1
Syn → 20% MOT17	51.7	48.0	56.3	62.1	54.4
Syn → 40% MOT17	57.4	54.3	61.5	69.3	61.3
Syn → 60% MOT17	59.7	57.1	63.0	70.0	62.9
Syn → 80% MOT17	60.6	58.1	63.8	73.0	67.1
Syn → 100% MOT17	**63.7**	**61.6**	**66.9**	**74.7**	**68.7**

Table 1. Tracking results on MOT17 with different amounts of real data.

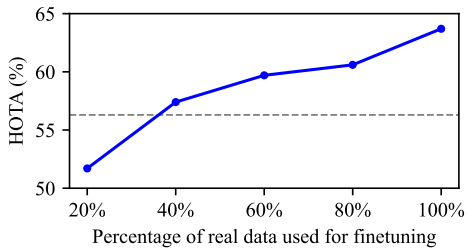

Figure 2. Plot showing tracking performances with different percentages of MOT17 data for fine-tuning. The dotted line represents the performance when the model is trained solely on 100% real data.

For the second research question, we extend the original MPT experiments on DanceTrack [22] in M^{3}Act by adjusting the synthetic data distributions. Specifically, we use the M^{3}Act data generator to prepare various datasets with different combinations of human groups. A total of 5 synthetic datasets are constructed, including 1K video clips of a single "Dance" group ("D"), 1.5K videos with a "walk" group and a "run" group simulated at the same time (denoted as "WalkRun" or "WR"), 2.5K videos of "WR"+"D", 6K videos of all "Single-group" activities, and 9K videos of all "Multi-group" activities (simulated simultaneously). We then investigate the influence of these synthetic data distributions on model performance in the downstream target dataset.

4.2. Benchmark Method

Recent studies [16, 25, 27, 28] have demonstrated the superior effectiveness of end-to-end methods over traditional tracking-by-detection approaches [2, 7, 24], particularly evident in challenging datasets like DanceTrack [22] and MOT [14, 19]. Therefore, our study aims to showcase the

Training Data	HOTA↑	DetA↑	AssA↑	IDF1↑	MOTA↑
DT	68.8	82.5	57.4	70.3	90.8
DT + Syn (D)	59.0	75.5	46.1	59.0	82.6
DT + Syn (WR)	70.1	83.1	59.4	72.5	92.0
DT + Syn (WR+D)	**71.9**	**83.6**	**62.0**	**74.7**	**92.6**
DT + Syn (Single-group)	65.1	80.2	55.8	66.7	89.1
DT + Syn (Multi-group)	**73.5**	**83.9**	**64.7**	**75.8**	**93.0**

Table 2. MPT results on DanceTrack under different synthetic data distributions (eg. group types). We use all single-group and multi-group data for the last two rows.

effectiveness of synthetic data in improving downstream task performance using end-to-end tracking models, which offer a better assessment of impact of synthetic data on overall tracking performance. We primarily evaluate the performance using MOTRv2 [28], the state-of-the-art method across various benchmark datasets. MOTRv2 is an extension of MOTR [27] by integrating YOLO-X [18] for enhanced detection bootstrapping.

4.3. Results

Tab. 1 presents the tracking results on MOT17 [19]. First, pretraining with our synthetic data leads to significant performance gain on all five metrics, compared to the model without pretraining. This observation aligns with several previous studies [3, 12, 23, 26] that adding synthetic data improves model performances. Second, the performance improves with the increased amount of real data, as illustrated in Fig. 2. The model finetuned on only 20% (or less) real data is inferior to the same model trained on 100% real data, which suggests a domain gap between the synthetic and real data. Last, the model trained solely on 100% real data achieves comparable performance to the model finetuned on 20% to 40% of real data. In other words, *our synthetic data is equivalent to 60% to 80% of MOT17 training data.*

We present in Tab. 2 the results on DanceTrack using various group types of synthetic data for training. These results reveal that training with multi-group synthetic data, such as "WR" and "Multi-group," leads to superior performance. The design of multi-group synthetic data, with multiple human activities animated in the same scene, provides frequent identity switches and produces high-complexity datasets that are suitable for challenging target datasets with dynamic movements such as DanceTrack. On the contrary, using synthetic data with each subject animated procedually and independently, such as "D" and "Single-group", yields inferior downstream performance to the baseline that is trained without any synthetic data. Our findings suggest that *synthetic datasets with apparent similarities in data complexity and distribution to target datasets tend to enhance model performance.* It also underscores the importance of having a flexible data generator, such as the modular and highly parameterized groups in `M³Act`, which allows for the customization of synthetic data distributions to closely approximate those of target real-world datasets.

4.4. Discussions

We discuss the following properties of synthetic data in our study and how the experimental results help to answer the research questions and provide insights for the potential implications of synthetic data.

4.4.1 Equivalency and Substitutability

Synthetic data has shown the potential to replace a substantial portion of real-world data, as evidenced by previous studies [3, 12]. For instance, results from [12] indicate that synthetic data generated by `M³Act` can substitute for 62.5% more real data from DanceTrack. In our study, we found that synthetic data can replace 60% to 80% of real-world data without sacrificing performance on MOT17. When considering the total number of image frames in both datasets, the equivalency ratio of synthetic to real data is approximately 30 times. Similarly, when considering the total number of track frames (or annotated bounding boxes), the ratio is roughly 5.6 times. These equivalency measurements provide valuable insights into the suitability and efficiency of different synthetic datasets for target downstream tasks. Furthermore, our results highlight that, despite existing domain gaps between synthetic and real datasets, the substitutability of synthetic data for target real-world data can significantly reduce data collection and annotation costs.

4.4.2 Impact of Synthetic Data Distributions

Although synthetic data offers practical benefits such as enhanced performance, substitutability, and cost reduction, its effectiveness relies on how closely its distribution matches that of the target real-world data. Our experiments suggest that not all synthetic datasets contribute equally to enhancing downstream task performances, even when designed for the same task. In fact, adding some synthetic data to the training dataset could even worsen performance. While most synthetic data generators [23, 26] offer limited parameterization options during the generation process, more complex tasks like multi-person tracking or scenarios involving multiple groups necessitate finer adjustments to the configuration parameters such as group types, person alignments, and group size within the data generator. Therefore, providing flexibility in synthetic data generators and adjusting parameters to generate synthetic data are crucial for narrowing domain gaps and ensuring the efficacy of using synthetic data in real-world applications.

5. Conclusion

In this study, we explore the equivalency, substitutability, and flexibility of synthetic data, offering insights into its practical applications in real-world scenarios. Despite our efforts and promising results shown in the experiments,

concrete evidence of complete substitution of synthetic data for real data in training datasets remains elusive. One significant challenge arises from the existence of domain gaps between synthetic and real data, which ultimately limit the adaptability of models trained solely on synthetic data. As we move forward, addressing these limitations will be crucial in maximizing the potential of synthetic data and enhancing its utility in various machine learning tasks.

6. Acknowledgement

The research was supported in part by NSF awards: IIS-1703883, IIS-1955404, IIS-1955365, RETTL-2119265, and EAGER-2122119. This work was also partially supported by the Center for Smart Streetscapes, an NSF Engineering Research Center, under cooperative agreement EEC-2133516. This material is based upon work supported by the U.S. Department of Homeland Security[1] under Grant Award Number 22STESE00001 01 01.

References

[1] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Hspace: Synthetic parametric humans animated in complex environments. *arXiv preprint arXiv:2112.12867*, 2021. 2

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 3

[3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1, 2, 3, 4

[4] Steve Borkman, Adam Crespi, Saurav Dhakad, Sujoy Ganguly, Jonathan Hogins, You-Cyuan Jhang, Mohsen Kamalzadeh, Bowen Li, Steven Leal, Pete Parisi, et al. Unity perception: Generate synthetic data for computer vision. *arXiv preprint arXiv:2107.04259*, 2021. 2

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[6] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3d human recovery. *arXiv preprint arXiv:2110.07588*, 2021. 2, 3

[7] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 3

[8] Che-Jui Chang. Transfer learning from monolingual asr to transcription-free cross-lingual voice conversion. *arXiv preprint arXiv:2009.14668*, 2020. 1

[9] Che-Jui Chang and Shyh-Kang Jeng. Acoustic anomaly detection using multilayer neural networks and semantic pointers. *Journal of Information Science & Engineering*, 37(1), 2021.

[10] Che-Jui Chang, Sen Zhang, and Mubbasir Kapadia. The ivi lab entry to the genea challenge 2022–a tacotron2 based method for co-speech gesture generation with locality-constraint attention mechanism. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 784–789, 2022.

[11] Che-Jui Chang, Long Zhao, Sen Zhang, and Mubbasir Kapadia. Disentangling audio content and emotion with adaptive instance normalization for expressive facial animation synthesis. *Computer Animation and Virtual Worlds*, 33(3-4): e2076, 2022. 1

[12] Che-Jui Chang, Danrui Li, Deep Patel, Parth Goel, Honglu Zhou, Seonghyeon Moon, Samuel S. Sohn, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. Learning from synthetic human group activities, 2023. 1, 2, 3, 4

[13] Che-Jui Chang, Samuel S Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. The importance of multimodal emotion conditioning and affect consistency for embodied conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 790–801, 2023. 1

[14] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 3

[15] Salehe Erfanian Ebadi, You-Cyuan Jhang, Alex Zook, Saurav Dhakad, Adam Crespi, Pete Parisi, Steven Borkman, Jonathan Hogins, and Sujoy Ganguly. Peoplesanspeople: a synthetic data generator for human-centric computer vision. *arXiv preprint arXiv:2112.09290*, 2021. 1

[16] Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9901–9910, 2023. 1, 3

[17] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees GM Snoek. Actor-transformers for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 839–848, 2020. 1

[18] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4

[19] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2, 3, 4

[20] Md Ashiqur Rahman, Jasorsi Ghosh, Hrishikesh Viswanath, Kamyar Azizzadenesheli, and Aniket Bera. Pacmo: Partner

[1] Disclaimer. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

dependent human motion generation in dyadic human activity using neural operators. *arXiv preprint arXiv:2211.16210*, 2022. 1

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[22] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022. 1, 3

[23] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 1, 2, 4

[24] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3

[25] Feng Yan, Weixin Luo, Yujie Zhong, Yiyang Gan, and Lin Ma. Bridging the gap between end-to-end and non-end-to-end multi-object tracking, 2023. 3

[26] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. *arXiv preprint arXiv:2303.17368*, 2023. 1, 2, 3, 4

[27] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022. 1, 3, 4

[28] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22056–22065, 2023. 1, 3, 4

[29] Honglu Zhou, Asim Kadav, Aviv Shamsian, Shijie Geng, Farley Lai, Long Zhao, Ting Liu, Mubbasir Kapadia, and Hans Peter Graf. Composer: Compositional reasoning of group activity in videos with keypoint-only modality. *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)*, 2022. 1