Synthetic-to-Real Adaptation for Complex Action Recognition in Surveillance Applications

Shuhong Lu[a], Zhangyu Jin[a], Vickram Rajendran[b], Michal Harari[b], Andrew Feng[a], and Celso M. De Melo[c]

[a]USC Institute for Creative Technologies, USA
[b]Applied Intuition, USA
[c]Army Research Laboratory, USA

ABSTRACT

In this paper, we propose to enhance action recognition accuracy by leveraging synthetic data and domain adaptation. Specifically, We achieve this through the creation of a synthetic dataset mimicking the Multi-View Extended Video with Activities (MEVA) dataset and the introduction of a multi-modal model for domain adaptation. This synthetic-to-real adaptation approach improves recognition accuracy by leveraging the synthetic data to enhance model generalization. Firstly, we focus on creating and utilizing synthetic datasets generated through a high-fidelity physically-based rendering system. The sensor simulation incorporates domain randomization and photo-realistic rendering to reduce the domain gap between the synthetic and real data, effectively addressing the persistent challenges of real data scarcity in action recognition.

Complementing the synthetic dataset generation, we leverage the multi-modal models in the synthetic-to-real adaptation experiments that utilize RGB images and skeleton features. Our experiments show that even relatively straightforward techniques, such as synthetic data pre-training, provide improvements to the models. Our work highlights the effectiveness of the approach and its practical applications across various domains, including surveillance systems, threat identification, and disaster response.

Keywords: synthetc data, action recognition, domain adaptation

1. INTRODUCTION

Action recognition plays a pivotal role in various applications such as surveillance systems, human-computer interaction, and robotics. In this research, we present a novel approach aimed at significantly enhancing the accuracy of action recognition models. Our method leverages synthetic data generation techniques coupled with domain adaptation strategies to address the challenge of data scarcity and improve model generalization. Recent advancements in action recognition have highlighted the importance of addressing domain shifts, where models trained on one dataset may not perform well when deployed in a different environment. To mitigate this issue, our approach integrates domain adaptation techniques to ensure robust performance across diverse scenarios.

Specifically, we propose the creation of a synthetic dataset closely resembling the MEVA dataset,[1] a widely used benchmark in action detection. Through this synthetic data augmentation, we enrich the training environments for existing models, particularly benefiting certain class labels with limited training samples. In the experiments, we simplify the target task into activity recognition by pre-processing the data to localize the activity in both spatial and temporal domain. We further processed the data with monocular human pose estimation model to predict the skeletal keypoints for each person. This allow us to leverage the multi-modal models with RGB images and skeleton features in the synthetic-to-real adaptation experiments. The models utilize a spatial-temporal transformer for RGB data and a graph-convolution network for skeletons. With the domain adaptation techniques on the multi-modal data, we demonstrated the superior generalization capabilities for models supplemented with synthetic data compared to models trained solely on real data. In summary, our experiments in synthetic-to-real adaptation reveal that our model outperforms those trained solely on real data, offering promising practical applications across diverse domains. These findings underscore the efficacy of our approach in addressing data scarcity issues while advancing the state-of-the-art in action recognition technology.

2. RELATED WORK

2.1 Group Activity Recognition

Recent advancements in group activity recognition have been propelled by novel approaches leveraging deep learning architectures and attention mechanisms. Tamura et al.[2] introduced a framework employing transformers for social group activity recognition, effectively capturing spatial and temporal dependencies among individuals within a group. Similarly, Yuan et al.[3] proposed a spatio-temporal dynamic inference network to model complex interactions among group members, achieving significant improvements in recognition accuracy. Duan et al.[4] revisited skeleton-based action recognition, enhancing the understanding of group dynamics through improved modeling of human skeletons. Furthermore, Zhou et al.[5] presented Composer, a compositional reasoning framework for group activity recognition in videos using only keypoint modalities, demonstrating the efficacy of their approach on diverse group activities. Zappardino et al.[6] addressed the challenge of learning group activities from skeletons without individual action labels, offering insights into unsupervised group activity recognition. Additionally, prior works such as the hierarchical long short-term concurrent memory by Shu et al.[7] and learning actor relation graphs by Wu et al.[8] have laid the groundwork for understanding complex human interactions and relations within groups. In our experiment, we leverage the recent findings to select strong baseline models that cover both RGB and skeletal keypoint modalities.

2.2 Domain Adaptation

Domain adaptation techniques have been extensively explored to improve the performance of computer vision models across different domains. Several studies have investigated the adaptation of models trained on source domain data to target domain distributions to enhance their generalization capabilities. Recent research in domain adaptation for various computer vision tasks has witnessed significant advancements leveraging deep learning techniques. Tsai et al.[9] proposed an adversarial learning method for domain adaptation in semantic segmentation tasks, achieving favorable results in various domain adaptation scenarios. Similarly, Chen et al.[10] addressed domain shift in object detection by introducing domain adaptive Faster R-CNN, effectively reducing the discrepancy between source and target domains at both image and instance levels. Additionally, Gupta et al.[11] explored the use of synthetic data for text localization in natural images, demonstrating its effectiveness in improving detection performance. Furthermore, recent works such as Wang et al.[12] and Chen et al.[13] introduced novel approaches for domain adaptation with attention mechanisms and progressive feature alignment, respectively, showcasing promising results across various tasks. While most of the previous research investigated domain adaptation for classification or detection tasks, relatively few works have focused on domain adaptation for human activity recognition. Munro and Damen[14] addressed domain adaptation for fine-grained action recognition, proposing a multi-modal approach combining self-supervision and adversarial alignment to mitigate domain shift in action recognition tasks. These methods collectively contribute to advancing the state-of-the-art in domain adaptation, offering effective solutions for handling domain shift across different visual recognition tasks. In this paper, we further explore the domain adaptation for activity recognition by creating a synthetic dataset that mimics real MEVA[1] data. This builds the foundation for investigating synthetic-to-real adaptation for group activity recognition task.

3. METHOD

3.1 Synthetic Data Generation

To augment the training data for activity recognition on the MEVA dataset, we leveraged the Unreal game engine to generate synthetic scenes and actions. The Unreal engine provides a powerful platform for creating highly realistic virtual environments and 3D avatar behaviors with dynamic lighting, physics, and interactions. By carefully designing scenes that mirror real-world scenarios present in the MEVA dataset, we can simulate various activities and actions performed by individuals or groups. This process involves creating 3D models of human characters, objects, and environments, as well as defining their animations and interactions. Additionally, we applied domain randomization by adding variations in lighting conditions, camera perspectives, and environmental factors to enhance the diversity of the synthetic data. Through this approach, we generate the MEVA-Syn dataset, a large number of annotated synthetic training data that closely resemble real-world

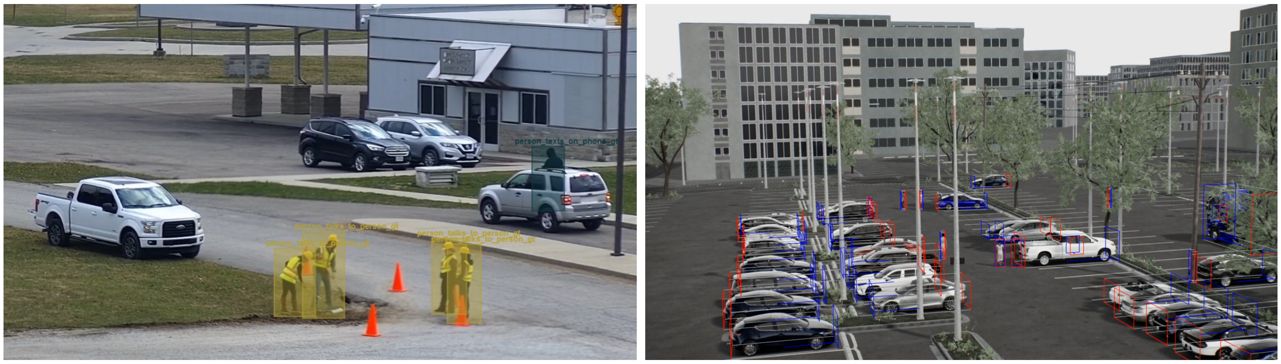
person_texts_on_phone_gt
person_talks_to_person_gt
person_talks_to_person_gt
person_talks_to_person_gt
person_talks_to_person_gt

Figure 1. Example frames of the dataset for our synthetic-to-real adaptation experiments. (Left) Real MEVA data (Right) our MEVA-Syn synthetic data.

scenarios captured in the MEVA dataset, thereby facilitating the training of robust action recognition models with improved generalization capabilities. Figure 1 shows an example frame from the real MEVA dataset and our MEVA-Syn synthetic dataset.

3.2 Action Recognition Architecture

To create a strong starting point for recognizing group activities, we draw from recent research of Dynamic Inference Network (DIN)[3] and COMPOSER[5] for RGB and skeleton modalities respectively. DIN models intricate interactions within group activities by forming person-specific interaction graphs for feature updates and global-level interaction fields with local initialization. On the other hand, COMPOSER reasons about group activities using only keypoints from videos. By breaking down group activities into compositional actions and reasoning over their spatial and temporal relationships, it achieves robust recognition without relying on pixel-level information. Our baseline models utilize both methods, leveraging DIN's dynamic modeling capabilities for RGB input and COMPOSER's compositional reasoning framework for skeletal keypoints. The experimentations with different modalities effectively captures complex spatiotemporal interactions and compositional structures within group activities, providing a strong baseline comparisons for group activity recognition.

3.3 Synthetic-to-Real Adaptation

We utilize a simple approach to domain adaptation that leverages both synthetic and real data to enhance model robustness and adaptability. Our method involves training the action recognition model and feature extractor using a combination of synthetic and real data initially. Subsequently, we fine-tune the model using only real data for synthetic-to-real adaptation, thereby aligning the model's representations with the target domain while preserving its ability to exploit synthetic data for improved generalization.

Initially, we train the action recognition model and associated feature extractor using the combined dataset that comprises both synthetic and real-world MEVA data. The synthetic data is generated using advanced rendering such as ray-tracing from Unreal Engine, which enables the creation of realistic and diverse action scenarios. By incorporating synthetic data during the initial training phase, the model learns to extract features and recognize actions across a wide range of environmental conditions, motion patterns, and lighting variations. After the initial training phase, we fine-tune the pre-trained model using labeled real-world data from the target domain. This fine-tuning process involves updating the model parameters based on the real data while preserving the learned representations from the synthetic data. By fine-tuning exclusively on real data, the model adapts its features and decision boundaries to better align with the characteristics of the target domain, thereby improving its performance on real-world action recognition tasks.

The key advantage of this approach lies in its ability to facilitate synthetic-to-real adaptation without sacrificing the benefits of synthetic data during initial training. By incorporating synthetic data during the initial training phase, the model learns rich and diverse feature representations that generalize well across domains. Subsequent fine-tuning using real data ensures that the model adapts its representations to the target domain

Datasets	Image Type	Multi-View	Multi-Group	Annotations							
				Bbox		Pose		Depth	Mask	Atomic Atn.	Group Act.
				2D	3D	2D	3D				
CAD	Real		✓	✓						✓	✓
Volleyball	Real		✓	✓						✓	✓
NTU-RGBD 120	Real	✓		✓			✓	✓		✓	
MEVA GAR	Real	✓	✓	✓						✓	✓
MEVA-Syn (Ours)	3D Scene	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. A comparison of synthetic datasets as well as commonly used real datasets for group activity understanding.

Datasets	GAR Category	# of Clips	Avg. Img. per Clip	Avg. Actor per Clip
MEVA GAR	person_talks_to_person	1,838	27.44	2.74
	person_enters_vehicle	495	24.92	1
	person_exits_vehicle	518	23.61	1
	person_talks_on_phone	207	23.05	1
	person_loads_vehicle	54	21.41	1
	person_unloads_vehicle	63	23.76	1
MEVA-Syn (Ours)	person_talks_to_person	103	74.85	2.34
	person_enters_vehicle	87	36.67	1
	person_exits_vehicle	172	39.95	1
	person_talks_on_phone	92	148.0	1
	person_loads_vehicle	71	143.76	1
	person_unloads_vehicle	103	148.0	1

Table 2. A summary of the real (MEVA GAR) and synthetic (MEVA-Syn) dataset used in our synthetic-to-real adaptation experiment. Note the imbalanced sample distribution in MEVA GAR dataset, which results in lower performance for long tail classes such as person_loads_vehicle and person_unload_vehicle.

while retaining the knowledge gained from synthetic data. As shown in the experiment section, this synthetic-to-real adaptation strategy effectively bridges the domain gap between synthetic and real-world action data, leading to improved performance and generalization capabilities in real-world scenarios.

4. EXPERIMENTS

In this section, we first present datasets and implementation details for Group Activity Recognition (GAR). Next, we showcase the practical utilities of our synthetic-to-real adaptation approach to GAR through two core experiments: Keypoint-based GAR, and RGB-based GAR.

4.1 Experiment Setup

MEVA GAR Dataset The Original MEVA[1] dataset is a large-scale dataset for human activity recognition, including 144 hours for 37 activity types, marking bounding boxes of actors and props. Next, we process it to complete the GAR task. Frames are extracted from raw MEVA videos with strides $[1, 5, 25, 125]$ given the length of the videos $[1, 50, 250, 1250, \infty]$. For each activity, it provides five kinds of annotations: i) starting and ending timestamp of the given activity; ii) coordinates of people's bounding boxes in the center frame; iii) identifications of the same person throughout the given clip; iv) group activity labels for the given clip v) individual action labels for the annotated person stay the same as group labels.

2D skeletons are generated by HRNet[15] following the COCO[16] format. Area of interest (ROI) is cropped at $H \times W = 480 \times 720$, where people included in the activity are placed at the center of ROI. For the experiment, we selected a subset of labels including *person_talks_to_person*, *person_enters_vehicle*, *person_exits_vehicle*, *person_talks_on_phone*, *person_loads_vehicle*, and *person_unloads_vehicle* and pre-processed the dataset from action

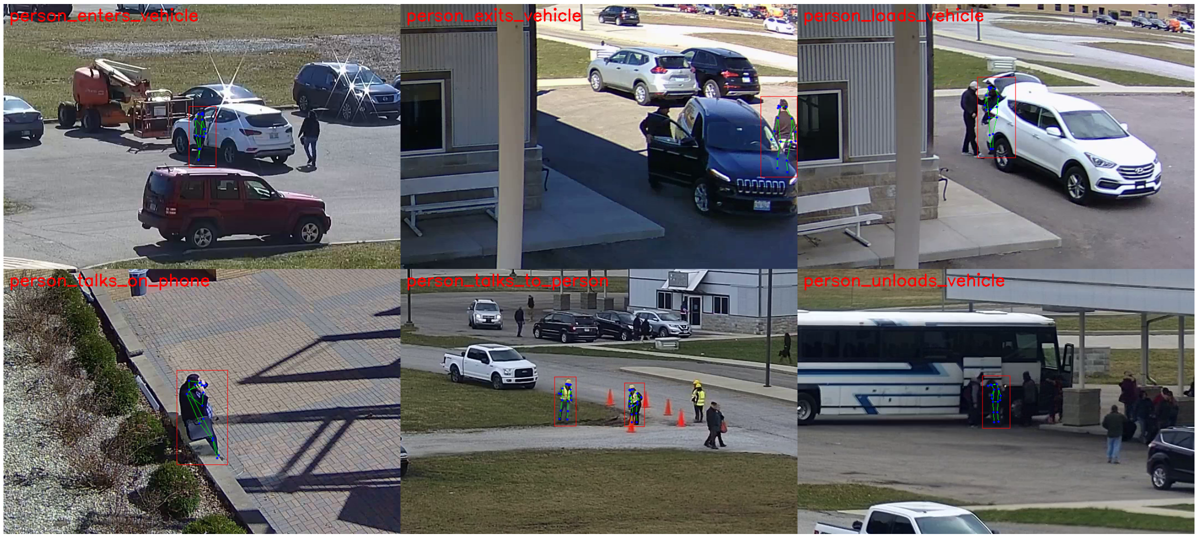

Figure 2. Example frames from real MEVA dataset showing the six activity categories used in the synthetic-to-real adaptation experiments.

detection into activity recognition form. Figure 2 shows the example frames of the selected activities. Class weights are set to $[0.009, 0.031, 0.028, 0.101, 0.274, 0.266]$ due to class imbalance shown in Table 2. Train-test split follows the original training and evaluation level annotations.[1] Two metrics are used for evaluating the performance of a model, i.e., MCA (%) which is short for Multi-class Classification Accuracy, and MPCA (%) which is short for Mean Per Class Accuracy.

Our MEVA-Syn Dataset is a multi-view multi-group multi-person human atomic action and group activity dataset specifically built to enhance the MEVA GAR dataset, as shown in Table 1, which comprises of 212 videos containing 150 frames. As discussed in Section 3.1, the dataset is created in Unreal using highly realistic virtual environments and 3D avatar behaviors to mimic the human activities recorded in the real MEVA dataset. Figure 3 shows the example frames of human activities following the motion and visual styles of the real MEVA dataset. For each activity, the GAR task will only include the following types of annotations: i) starting and ending timestamp of the given activity; ii) 3D keypoints of each person in absolute coordinates; iii) actor identifications and component identifications for each person; iv) ego vehicle's pose and camera intrinsic; v) group activity labels and person action labels with the same definition as the MEVA GAR dataset. 2D keypoints are then computed by projecting 3D keypoints to the camera view based on the ego's pose and camera intrinsic. As the initial annotated 2D bounding boxes are truncated upon occlusion, we employ the 2D projected keypoints' envelop bounding box as the ground truth 2D bounding box instead. Additionally, 2D skeletons based on HRNet are also provided to correspond with the keypoints definition and distribution in the MEVA GAR dataset. The ROI still remains the same at $H \times W = 480 \times 720$. Half of the videos are utilized for training and the other half for testing. Class weights are all set to 1 because of balanced categories shown in Table 2. Similarly, MCA (%) and MPCA (%) are used for evaluation.

Implementation Details The number of frames for each activity is fixed at 10. If the clip length is less than 10, the whole clip will be doubled until it has more than 10 frames. Otherwise, frames would be evenly selected throughout the whole clip to cover as much significant motion as possible. The number of people for each activity is fixed at 2. If the activity contains more than 2 participants, we would pick the first two instances. If not, the same individuals would be doubled in the scene. This is because most of the GAR categories, except *person_talks_to_person*, consist of only one actor, as shown in Table 2. Simple repetition avoids the model being overly replied on the number of actors. Additionally, following the adaptation method in Section 3.3, a two-stage strategy is applied during training time. The MEVA GAR dataset and our MEVA-Syn Dataset are combined together in Stage 1, while only the MEVA GAR dataset is included in Stage 2 to further fine-tune the model on real scenarios.

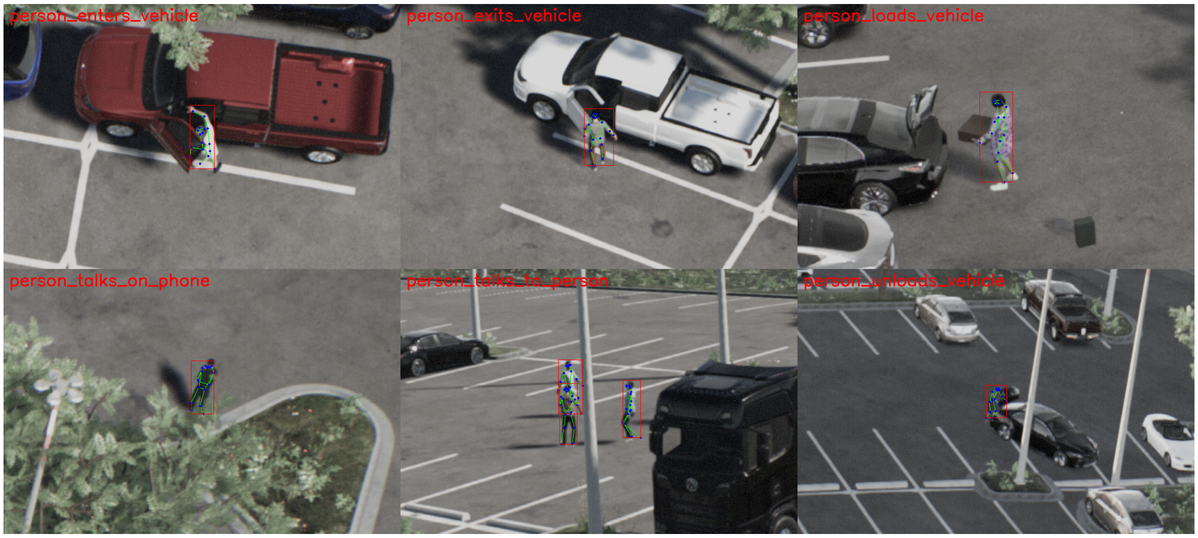

Figure 3. Example frames from real MEVA-Syn dataset demonstrating the corresponding behaviors created in the synthetic data to match the movements of human activities in real dataset.

4.2 Keypoint-based GAR

We consider Composer[5] as the benchmark model for the keypoint-only modality. It is a multi-scale Transformer-based architecture that performs attention-based reasoning over tokens at each scale and learns group activity compositionally. Input keypoints are normalized at three levels: i) absolute skeleton coordinates under image size; ii) relative skeleton Object Keypoint Similarity (OKS)[17] of the same person in the previous and current timestamp; iii) relative skeleton and bounding box coordinates at the same timestamp. Augmentations further enable Composer to obtain higher robustness, which includes Actor Dropout, Horizontal Flip, and Random Translation and Rotation. We use the person feature with embedding dimension $D = 1024$. We use online clustering with Sinkhorn iteration 3 and the number of clusters is set to 2. As for training settings, we apply batch size 384 and Adam optimizer with initial learning rate $5e-4$ and weight decay $1e-3$.

Keypoint-only Domain Adaptation Analysis Detailed MCA and MPCA can be found in Table 3. Unlike RGB-based GAR methods, keypoint-based ones suffer less from image style gaps between real and synthetic camera views. Instead, the motions of each person are more significant in the keypoint adaptation. And that is the reason why GAR categories with complicated motions do not benefit that much from adaptation, such as *person_talks_to_person*, *person_loads_vehicle*, and *person_unloads_vehicle*. On the other hand, motions with simple or fixed patterns gain performance in the adaptation, such as *person_enters_vehicle* and *person_exits_vehicle*.

Impact of Skeleton Generation Our MEVA-Syn dataset provides 2D keypoints annotations as described in the previous sections. Those keypoints are still visible even when the person is heavily occluded by vehicles or buildings, but the ones in the MEVA GAR dataset remain unstable when occlusion happens, see Figure 4. To check the impact of skeleton distribution, we perform ablation experiments between ground truth and estimated keypoints on the MEVA-Syn dataset. In Table 4, in both adaptation and non-adaptation circumstances, Composers[5] trained on HRNet's estimated keypoints perform better than those trained on ground truth keypoints.

MEVA GAR Analysis Keypoint-based GAR methods have some more limitations on the MEVA GAR dataset. In Figure 5, without the annotations of the objects in actors' hands, it would be hard to distinguish between *person_loads_vehicle* and *person_unloads_vehicle*. Because RGB pixels already have object information, the issue won't be as serious when using RGB-based techniques. Also, in Figure 6, keypoint-only recognition is hindered by situations such as actors who are too small or far away, truncated people by picture edges, and mismatches between keypoints and bounding boxes. Overall, keypoint-based methods are not sensitive under image styles, but they suffer more in those mentioned hard cases.

Model	Train		Test		MCA	MPCA	Recall per class % ↑					
	MEVA	MEVA	MEVA	MEVA	% ↑	% ↑	talks_to	enters	exits	talks_on	loads	unloads
	GAR	-Syn	GAR	-Syn			person	vehicle	vehicle	phone	vehicle	vehicle
Composer (Keypoint)	✓		✓		85.34	**59.01**	**97.29**	**72.53**	74.89	69.64	**29.27**	**10.45**
		✓		✓	86.30	86.19	94.5	76.32	82.82	87.36	86.96	89.22
		✓	✓		40.39	26.55	46.0	27.47	39.95	43.45	2.44	0
	✓	✓	✓		**86.05**	58.22	96.93	65.3	**82.42**	**90.48**	9.76	4.48
DIN (RGB)	✓		✓		68.10	29.03	91.54	14.06	**63.55**	**5.05**	0	0
		✓		✓	69.46	68.39	77.36	56.72	84.06	41.9	73.25	77.08
	✓	✓	✓		**71.18**	**30.26**	**95.3**	**42.4**	40.53	3.37	0	0

Table 3. Experiment results for synthetic-to-real adaptation using RGB and skeleton keypoint modalities. Combining MEVA-Syn and MEVA GAR produces best results in MCA when tested with real MEVA data for both modalities.

Model	Skeleton		Train		MCA	MPCA	Recall per class % ↑					
	GT	HRNet	MEVA	MEVA	% ↑	% ↑	talks_to	enters	exits	talks_on	loads	unloads
			GAR	-Syn			person	vehicle	vehicle	phone	vehicle	vehicle
Composer	✓			✓	29.67	18.17	35.22	15.9	34.02	12.5	**2.44**	**8.96**
		✓		✓	**40.39**	**26.55**	**46.0**	**27.47**	**39.95**	**43.45**	2.44	0
	✓		✓	✓	85.51	**60.43**	96.39	**73.73**	73.06	85.12	**26.83**	**7.46**
		✓	✓	✓	**86.05**	58.22	**96.93**	65.3	**82.42**	**90.48**	9.76	4.48

Table 4. Comparison of synthetic-to-real adaptation results using ground truth skeleton keypoint from the synthetic data and estimated keypoint from HRNet. The keypoints extracted from HRNet has smaller domain gap to real data, and produces better test result when training with only MEVA-Syn data. The HRNet keypoint also performs slightly better when combining MEVA real data and MEVA-Syn in the training.

4.3 RGB-based GAR

We utilize DIN[3] as the baseline model for RGB-based GAR experiment. We trained the model using SGD optimizer with learning rate 10^{-4}. During training, RandomRescale, RandomCrop and Horizontal Flip are applied as data augmentation to improve model robustness. As shown in Tabel 3, the experiment results with RGB-only modality show improvements in both MPCA and MPCA metrics when combining MEVA-Syn and real MEVA data. Overall, the experiment results support the benefit of supplementing the training of recognition model with synthetic data. For the per-label accuracy, the results indicated consistent findings that the model has difficulty adapting for long-tail categories such as person_loads_vehicle and person_unloads_vehicle. The results also showed that the RGB modality perform worse in generalization than skeletal keypoints in both real data and synthetic-to-real adaptation. This is evident in accuracy gap between COMPOSER and DIN in real MEVA data benchmark using both real and synthetic training data. The main reason for this performance gap might be due to the training dataset size. With a limited number of training samples, RGB model tends to perform worse as it requires more samples to cover appearance variations from the environment and persons in addition to human motion variations. On the other hand, keypoint-based model only needs to account for variations in motion behaviors since the input data contain only keypoint locations and are invariant to changes in appearances.

5. CONCLUSION AND FUTURE WORKS

In this paper, we have aimed at enhancing action recognition accuracy in surveillance applications through the integration of synthetic data and domain adaptation techniques. This is done by leveraging synthetic datasets closely resembling the MEVA dataset and conduct experiemtns with multi-modal models for domain adaptation. Through the creation and utilization of synthetic datasets generated using a high-fidelity rendering system, coupled with domain randomization techniques, we have effectively mitigated challenges arising from real data scarcity in action recognition. For future works, we hope to develop the multi-modal model that integrate RGB images and skeleton features to find a synergistic fusion of spatial and temporal features.

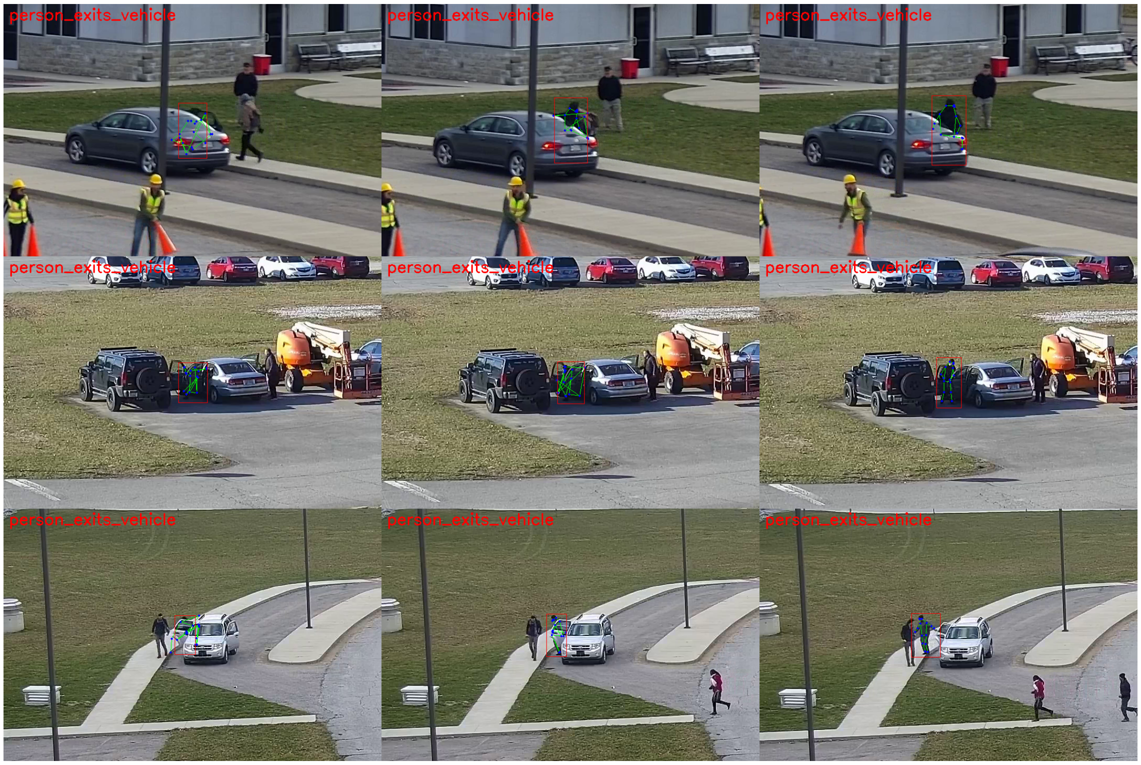

Figure 4. Example frames demonstrating the occlusion problems in MEVA GAR dataset for estimating keypoint with HRNet. In the occluded regions, the predicted keypoints are instable and sometimes not able to form a coherent human skeleton.

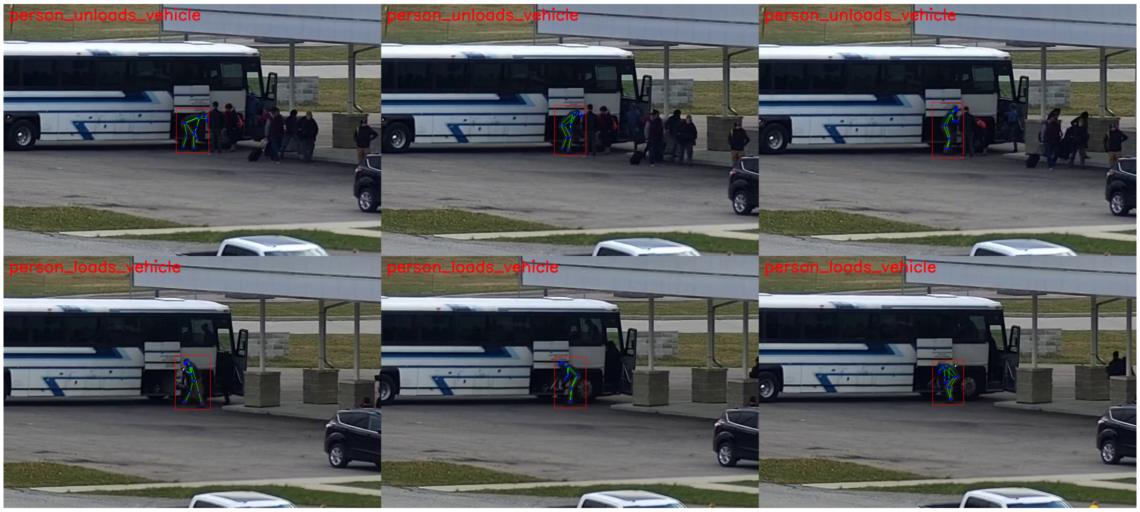

Figure 5. Example frames showing the difficulty for learning load & unload behaviors using keypoint modality. Since there is not enough context about the objects from only the skeletal keypoint to determine the actual action, it is confusing for model to distinguish between the two behaviors.

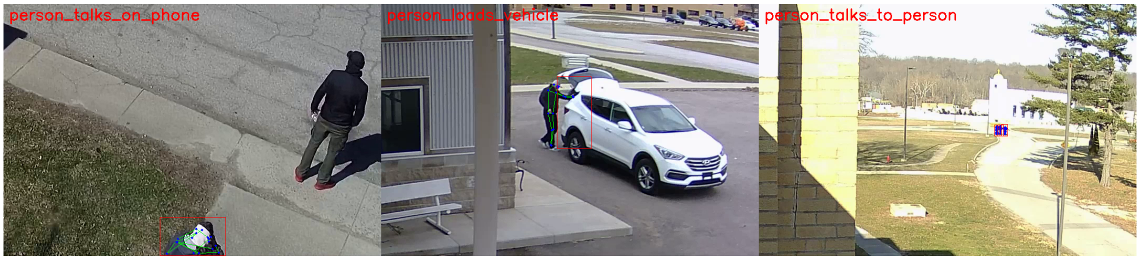

Figure 6. Example frames demonstrating the difficult cases in MEVA GAR dataset. Truncated (Left), Bbox & Skeleton Mismatch (Middle), Too far too small (Right)

Our experiments demonstrate the efficacy of our approach in improving action recognition accuracy, even with relatively straightforward techniques such as synthetic data pre-training. By integrating synthetic data during initial training and employing domain adaptation techniques for fine-tuning on real data, our model showcases superior generalization capabilities compared to models trained solely on real data. These findings underscore the practical applications of this approach for action recognition problems. In the future, we would like to expand the synthetic data generations with more activity categories and motion variations to better model the real-world human behaviors. We also hope to extend the experiments by exploring recent feature alignment methods for domain adaptation. Since many such methods were developed for static tasks such as object detection or semantic segmentation, extending such methods for temporal data will be a valuable research direction for activity recognition.

ACKNOWLEDGMENTS

The project or effort depicted was or is sponsored by by the DEVCOM Army Research Lab under contract number W911QX-21-D-0001. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

[1] Corona, K., Osterdahl, K., Collins, R., and Hoogs, A., "Meva: A large-scale multiview, multimodal video dataset for activity detection," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 1060–1068 (January 2021).

[2] Tamura, M., Vishwakarma, R., and Vennelakanti, R., "Hunting group clues with transformers for social group activity recognition," in [*Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*], 19–35, Springer-Verlag, Berlin, Heidelberg (2022).

[3] Yuan, H., Ni, D., and Wang, M., "Spatio-temporal dynamic inference network for group activity recognition," in [*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*], 7456–7465, IEEE Computer Society, Los Alamitos, CA, USA (oct 2021).

[4] Duan, H., Zhao, Y., Chen, K., Lin, D., and Dai, B., "Revisiting skeleton-based action recognition," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 2969–2978 (2022).

[5] Zhou, H., Kadav, A., Shamsian, A., Geng, S., Lai, F., Zhao, L., Liu, T., Kapadia, M., and Graf, H. P., "Composer: Compositional reasoning of group activity in videos with keypoint-only modality," in [*European Conference on Computer Vision*], 249–266, Springer (2022).

[6] Zappardino, F., Uricchio, T., Seidenari, L., and Del Bimbo, A., "Learning group activities from skeletons without individual action labels," in [*2020 25th International Conference on Pattern Recognition (ICPR)*], 10412–10417, IEEE (2021).

[7] Shu, X., Tang, J., Qi, G.-J., Liu, W., and Yang, J., "Hierarchical long short-term concurrent memory for human interaction recognition," *IEEE transactions on pattern analysis and machine intelligence* **43**(3), 1110–1118 (2019).

[8] Wu, J., Wang, L., Wang, L., Guo, J., and Wu, G., "Learning actor relation graphs for group activity recognition," in [*Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*], 9964–9974 (2019).

[9] Tsai, Y., Hung, W., Schulter, S., Sohn, K., Yang, M., and Chandraker, M., "Learning to adapt structured output space for semantic segmentation," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 7472–7481, IEEE Computer Society, Los Alamitos, CA, USA (jun 2018).

[10] Chen, Y., Li, W., Sakaridis, C., Dai, D., and Gool, L. V., "Domain adaptive faster r-cnn for object detection in the wild," in [*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 3339–3348, IEEE Computer Society, Los Alamitos, CA, USA (jun 2018).

[11] Gupta, A., Vedaldi, A., and Zisserman, A., "Synthetic data for text localisation in natural images," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 2315–2324, IEEE Computer Society, Los Alamitos, CA, USA (jun 2016).

[12] Wang, X., Li, L., Ye, W., Long, M., and Wang, J., "Transferable attention for domain adaptation," in [*Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*], *AAAI'19/IAAI'19/EAAI'19*, AAAI Press (2019).

[13] Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T., and Huang, J., "Progressive feature alignment for unsupervised domain adaptation," in [*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 627–636, IEEE Computer Society, Los Alamitos, CA, USA (jun 2019).

[14] Munro, J. and Damen, D., "Multi-modal domain adaptation for fine-grained action recognition," in [*2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*], 119–129, IEEE Computer Society, Los Alamitos, CA, USA (jun 2020).

[15] Sun, K., Xiao, B., Liu, D., and Wang, J., "Deep high-resolution representation learning for human pose estimation," in [*CVPR*], (2019).

[16] Lin, T.-Y., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L., "Microsoft coco: Common objects in context.," *CoRR* **abs/1405.0312** (2014).

[17] Maji, D., Nagori, S., Mathew, M., and Poddar, D., "Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 2637–2646 (2022).