Cross-Modal Knowledge Distillation for Human Trajectory Prediction in Virtual Reality

Franz Franco Gallo, Hui-Yin Wu, Lucile Sassatelli

Cross-Modal Knowledge Distillation for Human Trajectory Prediction in Virtual Reality

Franz Franco Gallo[1], Hui-Yin Wu[1], and Lucile Sassatelli[2,3]

[1] Université Côte d'Azur, Inria, Sophia-Antipolis, France
[2] Université Côte d'Azur, CNRS, I3S, Sophia-Antipolis, France
[3] Institut Universitaire de France, Paris, France
franz.franco-gallo@inria.fr

Abstract. Scene context informing on spatio-temporal interactions between people and other entities significantly improves accuracy of activity recognition and motion forecasting tasks, such as human trajectory prediction, but is difficult to obtain. Virtual reality (VR) offers an opportunity to generate and simulate diverse scenes with contextual information, which can potentially inform real-life scenarios. We design a teacher model leveraging heterogeneous graphs constructed from VR scene annotations to enhance prediction accuracy. This ongoing work proposes cross-modal knowledge distillation (CMKD), transferring the knowledge from the VR-constructed graphs to a student model that uses scene point clouds. Preliminary results show the potential of CMKD to transfer contextual information that significantly improves the prediction accuracy of the student model. Scene context informing on spatio-temporal interactions between people and other entities significantly improves accuracy of activity recognition and motion forecasting tasks, such as human trajectory prediction, but is difficult to obtain.

Keywords: Virtual Reality · Knowledge distillation · Cross-modal · Human trajectory prediction · Heterogeneous graphs

1 Introduction

Human trajectory prediction is essential for urban applications like autonomous driving and anomaly detection. Contextual information such as where pedestrians look, or the color of the traffic light, captures dynamic interactions and spatial-temporal relationships, thus improving trajectory prediction. However, obtaining rich context in real-world settings is challenging. Knowledge distillation, which allows the transfer of knowledge from a teacher model to a student one, can help the student model infer high-level context (e.g., activities, interactions) from a low-level presentation (e.g., pixels, voxels). Zhao et al. [3] successfully applied cross-modal knowledge distillation (CMKD) between a LiDAR-camera teacher model to a radar-camera student model.

We investigate CMKD on human trajectory prediction from a teacher model that has access to VR-constructed heterogeneous graphs to a student model that

operates only on scene point clouds that can be obtained from LiDAR sensors or photogrammetry. We use the CREATTIVE3D dataset [2] [1], which includes data from 40 participants walking in virtual street scenes with various motion and attention capture modalities and scene annotations. The teacher model integrates motion, gaze, and heterogeneous graphs, employing Graph Convolutional Networks (GCN) and cross-modal attention mechanisms to capture both static and dynamic elements of the scene. By enabling the estimation of graph interactions from unstructured point cloud data, the student model infers transferable features without additional annotations. Our method bridges the gap between the rich context data in VR and the unstructured data from real-world sensors, showing the advantages of immersive VR environments for in-context motion prediction tasks, which have the potential to inform on other tasks for human behaviour understanding.

2 Methodology

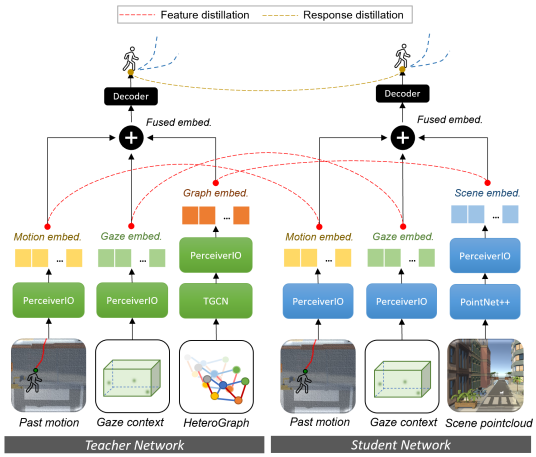


Fig. 1: Overview of our Cross-modal Knowledge Distillation Framework

We use the CREATTIVE3D dataset collected in six VR urban scenarios [2], which includes gaze and motion data of users carrying out road crossing tasks, and annotations that can be used to generate dynamic scene graphs, as shown in Fig. 1. Users carry out diverse and complex interactions within a virtual urban environment, such as interacting with traffic lights and transporting objects, which is difficult to identify in real-world settings. We evaluate the human trajectory prediction task on the 2D head position given past positions and contextual data. At a given time t (in frames), the human model comprises the head position $\mathbf{p}_t \in \mathbf{R}^2$ in meters, representing the user's absolute position. The model predicts a motion sequence over a horizon H, defined as

[1] https://zenodo.org/doi/10.5281/zenodo.8269108

$\mathbf{M}_{t+1:t+H} = \{(\hat{\mathbf{p}}_{t+1}), \ldots, (\hat{\mathbf{p}}_{t+H})\}$ from t, given past motion $\mathbf{M}_{t-x:t}$. In our models we use $x = 6$ and $H = 10$.

The dataset consists of 1,038 samples (832K frames), divided into 70% for training, 20% for validation, and 10% for testing. Performance is evaluated using the Average Displacement Error (ADE) and Final Displacement Error (FDE), which are the distance between the predicted and actual trajectory, at each frame and at the endpoint respectively.

2.1 Cross-Modal Knowledge Distillation Framework

Teacher Model: The teacher model takes as input the user's past position, gaze, and heterogeneous scene graph encoded with a Temporal Graph Convolutional Network (TGCN). We use PerceiverIO modules to fuse the multiple modalities and then decode them into the predicted trajectory [4].

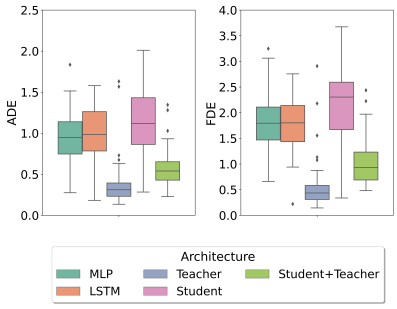
Student Model: The student model takes as input the user's past position, gaze, and the scene point cloud, which we use PointNet++ to extract meaningful features. These features are then resized using an adaptation layer, which adjusts the dimensions of the features to match the size of the scene graph embedding for accurate distillation loss computation. Finally, the PerceiverIO architecture fuses these modalities and decodes them in a manner similar to the teacher model.

Training: Our CMKD framework involves two-stages of training: (1) the teacher model is pre-trained on the dataset, and (2) the student model is trained on the same dataset with the soft targets (predicted trajectory) generated by the teacher model. The student model thus learns to replicate the teacher's predictions, leveraging the context from the TGCN embeddings.

We use both intermediate and response features for distillation. For intermediate features, the teacher's motion, gaze, and graph embeddings are matched with the student's corresponding feature embeddings using L2 loss [3]. For response features, the actual motion predictions of both models are compared using L1 loss. This approach ensures that the student model mimics the teacher's performance, improving trajectory predictions on the sensor-derived data. The total loss used in our training is given by $L_{\text{total}} = \alpha L_{\text{intermediate}} + (1-\alpha) L_{\text{response}}$ where $\alpha(0.5)$ is a weighting factor that balances the contribution of intermediate feature alignment and response prediction accuracy.

3 Results

We compare the performance of our student model against two baseline models: MLP and LSTM. The MLP model serves as a simple baseline that uses past positions to predict future trajectories without incorporating contextual information. The LSTM modelis based on TRACK [1] and fuses gaze data with motion data.



Architecture	ADE	FDE
MLP	0.959	1.835
LSTM	1.021	1.784
Teacher	0.381	0.579
Student w/o teacher	1.155	2.133
Student w/ teacher	0.581	1.023

Fig. 2: Comparison of ADE and FDE for different architectures. (Left) Boxplots. (Right) Table with mean values. The architectures compared include the baseline MLP, LSTM, Teacher model, Student model without teacher, and Student model with teacher.

Results in Fig. 2 show that the student model, trained with knowledge distilled from the teacher, achieves substantial improvements over the baselines. ADE and FDE are reduced by 49.7% and 52.0% respectively on the test set compared to the student model trained without the teacher. This demonstrates the efficacy of cross-modal distillation in enhancing trajectory prediction accuracy using more accessible point cloud data.

4 Conclusion

This ongoing work explores context-based human trajectory prediction through cross-modal knowledge distillation. By distilling knowledge from heterogeneous graphs constructed in VR to scene point clouds, the student model inherits the contextual understanding encoded by the teacher model. This transfer of knowledge allows the student model to make more informed predictions, reflecting a deeper comprehension of the scene's dynamics and interactions.

Future work will focus on refining the distillation process, exploring additional modalities, and enhancing the student's contextual understanding. This approach shows promise for improving trajectory prediction accuracy and enabling practical applications of VR and the metaverse for potential real-world applications.

Acknowledgements This work has been partially supported by the French National Research Agency through the ANR CREATTIVE3D project ANR-21-CE33-0001 and UCA^{JEDI} Investissements d'Avenir ANR-15-IDEX-01 (IDEX reference center for extended reality XR^{2}C^2). This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011014115R1 made by GENCI.

References

1. Rondón, M.F.R., Sassatelli, L., Aparicio-Pardo, R., Precioso, F.: Track: A new method from a re-examination of deep architectures for head motion prediction in 360^o videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5681–5699 (2021)
2. Wu, H.Y., Robert, F.A.S., Gallo, F.F., Pirkovets, K., Quere, C., Delachambre, J., Ramanoël, S., Gros, A., Winckler, M., Sassatelli, L., Hayotte, M., Menin, A., Kornprobst, P.: Exploring, walking, and interacting in virtual reality with simulated low vision: a living contextual dataset (2023), https://inria.hal.science/hal-04429351, preprint
3. Zhao, L., Song, J., Skinner, K.A.: Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15470–15480 (June 2024)
4. Zheng, Y., Yang, Y., Mo, K., Li, J., Yu, T., Liu, Y., Liu, C.K., Guibas, L.J.: Gimo: Gaze-informed human motion prediction in context. In: European Conference on Computer Vision. pp. 676–694. Springer (2022)