This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in: László Kopácsi, Michael Barz, Omair S. Bhatti, Daniel Sonntag: IMETA: An Interactive Mobile Eye Tracking Annotation Method for Semi-Automatic Fixation-to-AOI Mapping. Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, Association for Computing Machinery 2023. https://doi.org/10.1145/3581754.3584125

IMETA: An Interactive Mobile Eye Tracking Annotation Method for Semi-automatic Fixation-to-AOI mapping

László Kopácsi
laszlo.kopacsi@dfki.de
German Research Centre for Artificial Intelligence
Saarbrücken, Saarland, Germany

Michael Barz
michael.barz@dfki.de
German Research Centre for Artificial Intelligence
Saarbrücken, Germany
University of Oldenburg
Oldenburg, Germany

Omair Bhatti
omair_shahzad.bhatti@dfki.de
German Research Centre for Artificial Intelligence
Saarbrücken, Germany

Daniel Sonntag
daniel.sonntag@dfki.de
German Research Centre for Artificial Intelligence
Saarbrücken, Germany
University of Oldenburg
Oldenburg, Germany

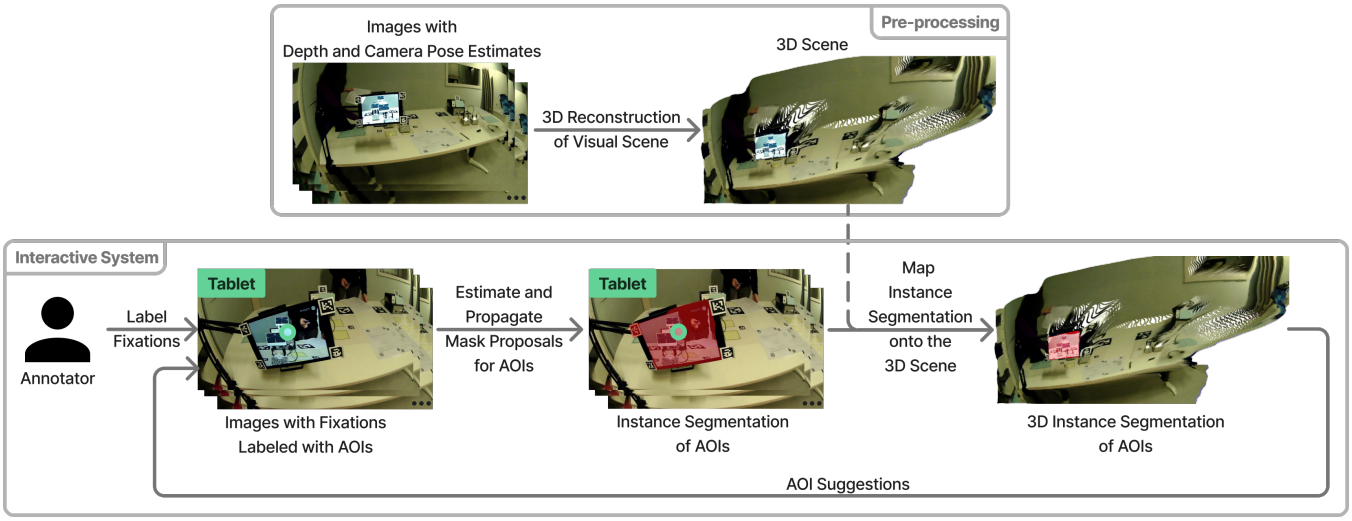

Figure 1: High-level overview of the proposed system. In this example, the 3D scene is reconstructed using a monocular depth prediction method called AdelaiDepth [35], and the video object segmentation is performed using XMem with f-BRS [5, 25].

ABSTRACT

Mobile eye tracking studies involve analyzing areas of interest (AOIs) and visual attention to these AOIs to understand how people process visual information. However, accurately annotating the data collected for user studies can be a challenging and time-consuming task. Current approaches for automatically or semi-automatically analyzing head-mounted eye tracking data in mobile eye tracking studies have limitations, including a lack of annotation flexibility or the inability to adapt to specific target domains. To address this problem, we present IMETA, an architecture for semi-automatic fixation-to-AOI mapping. When an annotator assigns an AOI label to a sequence of frames based on the respective fixation points, an interactive video object segmentation method is used to estimate the mask proposal of the AOI. Then, we use the 3D reconstruction of the visual scene created from the eye tracking video to map these AOI masks to 3D. The resulting 3D segmentation of the AOI can be used to suggest labels for the rest of the video, with the suggestions becoming increasingly accurate as more samples are provided by an annotator using interactive machine learning (IML). IMETA has the potential to reduce the annotation workload and speed up the evaluation of mobile eye tracking studies.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; Empirical studies in HCI; • **Computing methodologies** → *Machine learning*.

KEYWORDS

mobile eye tracking, areas of interest, annotation, interactive machine learning, 3D reconstruction, video object segmentation, fixation to aoi mapping

1 INTRODUCTION

The annotation of objects in videos is an important task for analyzing visual scenes and understanding the interactions and relationships between objects in a scene. It also provides training data for machine learning models, which can be used to improve object recognition and scene understanding tasks. However, traditional annotation methods require extensive manual segmentation of objects, which can be time-consuming and error-prone. Mobile eye tracking studies, which involve the use of head-mounted eye tracking devices to measure visual attention to specific areas of interest (AOIs) in a scene, present additional challenges for annotation due to the uniqueness of the scene videos for each participant and the potential for sudden movements and unusual viewpoints of the AOIs. Accurate annotation of mobile eye tracking data is therefore a challenging and time-consuming task.

Current approaches for annotating AOIs in mobile eye tracking data include manual annotation by one or more annotators [15, 30], the use of fiducial markers attached to target stimuli [2, 17, 21, 36], or the use of computer vision models [3, 7, 8, 13, 14, 16, 20, 22, 26, 28–30, 33]. However, each of these approaches has its own limitations. Manual annotation is tedious and susceptible to errors, the use of fiducial markers is obtrusive and not suitable for uninstrumented environments, and computer vision models may be limited by their reliance on pre-trained models that cannot be adapted to a specific target domain or a lack of flexibility with no possibility to adapt the model during the annotation process [3, 7, 8, 13, 14, 16, 20, 22, 26, 28–30, 33].

To address these limitations, we propose a novel interactive machine learning approach for the semi-automatic annotation of AOIs in mobile eye tracking data using a combination of interactive video object segmentation and 3D reconstruction. A high-level overview of how the proposed interaction system is expected to work is shown in Figure 1. The eye tracking recordings displayed in the figure are from a mobile eye tracking dataset collected at Saarland University for the purpose of investigating the effects of augmented reality (AR) support in a laboratory-based learning scenario [1, preregsitered at Open Science Framework].

2 RELATED WORK

Interactive video object segmentation is a method for segmenting and tracking objects of interest in a video. It involves using user input to guide the segmentation process and improve the accuracy of the resulting object masks. In the context of semi-automatic fixation-to-AOI mapping in mobile eye tracking studies, interactive video object segmentation could be used to efficiently annotate AOIs in the video by allowing the annotator to provide labels for a sequence of frames marked with fixation positions, and using the fixations to estimate and propagate the mask proposal of the AOI.

However, interactive video object segmentation has its limitations. It can struggle to re-identify highly similar objects, and can be sensitive to sudden camera movements, in such cases requiring a large amount of user input [5, 6, 11, 31]. It may also only provide reliable results under certain constraints, such as when the camera is static [9, 12]. To overcome these limitations, we suggest using 3D scene reconstruction in addition to interactive video object segmentation.

3D scene reconstruction is a process for creating a 3D model of a real-world environment or scene from images or video. It involves extracting geometric and semantic information from the images or video and using this information to build a 3D representation of the scene. There are various approaches to reconstruct 3D scenes from posed RGB-D images [18], some that only require the existence of depth or posed images besides RGB [19, 24, 34, 37], and a handful that can work with monocular RGB videos only [4, 10, 27, 32].

3D scene reconstruction can address the issues with interactive video object segmentation, such as re-identification, unusual viewpoints, and spatial and temporal consistency. By reconstructing the visual scene in 3D, we can map the AOI segments to 3D, minimizing the impacts of these problems by splitting the video into shorter segmentation tasks and aggregating the predicted AOI segments in the 3D scene. The resulting 3D segmentation of the AOI can then provide label suggestions for the rest of the video, streamlining the annotation process. We propose an IML framework that combines these technologies to tackle the limitations of current approaches for annotating AOIs in mobile eye tracking videos.

3 INTERACTIVE MOBILE EYE TRACKING ANNOTATION

Interactive Mobile Eye Tracking Annotation (IMETA) is a novel approach for semi-automatic fixation-to-AOI mapping in videos recorded by mobile eye trackers. The proposed architecture of IMETA is shown in Figure 2. A web interface shall serve as a user interface for IMETA and the logic itself is planned to be implemented as an interactive machine learning service.

The method begins with a pre-processing step to reconstruct the 3D scene from the mobile eye tracker recordings. This step requires camera poses and depth information for each frame, which can be estimated using monocular scene reconstruction methods. If this information is already recorded by the mobile eye tracker, this step can be skipped.

Through the frontend, the annotator can label fixations and correct AOI suggestions. The annotator is presented with an image marked with fixation positions. When a new fixation is labeled, a state-of-the-art interactive video object segmentation method, called XMem [5], estimates the mask proposal for the fixation, and propagates and aggregates the masks for the duration of the fixation. These segments are then mapped onto the 3D scene to create a 3D instance segmentation of the annotated AOI (i.e., the attended object). By mapping the fixation positions onto the 3D AOI segmentation and checking their corresponding label, we can automatically label the rest of the video. The 3D segmentation can be further refined by confirming or correcting label suggestions, and by labeling additional fixations. Based on the available annotations, the accuracy of the system is calculated, and the annotation can be stopped when the desired threshold is reached. We expect that by annotating only a handful of fixation sequences, the proposed system can automatically determine the AOI labels for the rest of the video. Furthermore, given a static environment, a single annotated video could generalize for all recordings from the same environment.

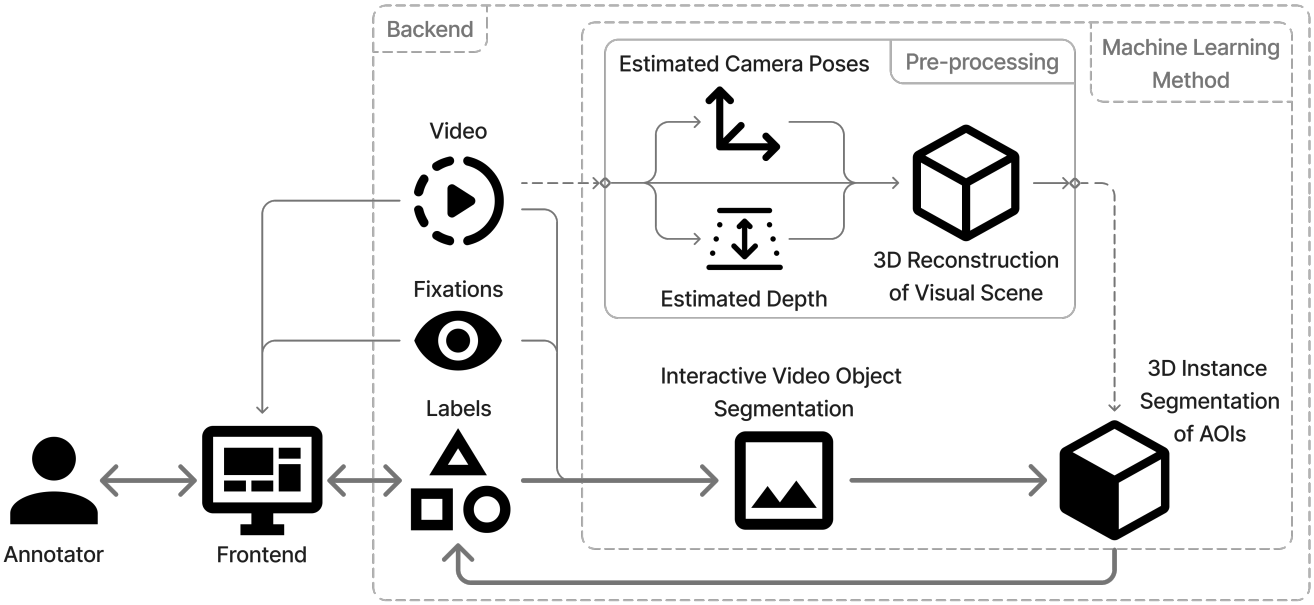

Figure 2: Proposed architecture of IMETA. It consists of a pre-processing step to reconstruct the 3D scene from the mobile eye tracker recording, an interface for interactive fixation-to-AOI mapping, an interactive video object segmentation method to estimate the mask proposal of the AOIs, and a module to map the AOI segments onto the 3D scene. The resulting 3D instance segmentation of the AOIs can then be used to generate label suggestions for each fixation. The arrows indicate the flow of data through the different components of the architecture.

4 CONCLUSION

In this paper, we proposed IMETA, a novel system for semi-automatic fixation-to-AOI mapping in videos recorded by mobile eye trackers. IMETA combines monocular 3D scene reconstruction with interactive video object segmentation to annotate AOIs. By utilizing 3D reconstruction, IMETA aims to overcome the challenges of traditional video annotation methods, such as the need for re-identification, handling unusual viewpoints, and ensuring spatial and temporal consistency. We believe that IMETA has the potential to greatly reduce annotation workload for evaluating studies that use mobile eye tracking videos.

However, the reliance on monocular 3D scene reconstruction methods may be a potential limitation, as it may not always provide accurate camera pose and depth estimates. Our future plans include using it with an AR headsets that include eye tracking capabilities or with head-mounted eye trackers equipped with RGB-D cameras, to overcome the limitations of monocular 3D reconstruction methods. The 3D reconstruction and segmentation of the visual scene could be done in real-time by the experimenter using the headset. Further, 3D scene reconstruction is typically constrained to static environments. We plan to integrate a solution to enable dynamic environment support similar to Kimera [23]. IMETA could then serve as an interactive annotation tool for annotating AOIs in mobile eye tracking studies on-the-fly. We plan to evaluate the performance of IMETA through experiments and user studies.

ACKNOWLEDGMENTS

This work was funded by the German Federal Ministry of Education and Research under grant number 01JD1811C (GeAR) and by the European Commission project MASTER (grant number 101093079; https://www.master-xr.eu/).

REFERENCES

[1] Kristin Altmeyer, Sebastian Kapp, Michael Barz, Luisa Lauer, Sarah Malone, Jochen Kuhn, and Roland Brünken. 2020. The effect of augmented reality on global coherence formation processes during STEM laboratory work in elementary school children. (Oct. 2020). https://doi.org/10.17605/osf.io/gwhu5

[2] Michael Barz, Florian Daiber, Daniel Sonntag, and Andreas Bulling. 2018. Error-aware gaze-based interfaces for robust mobile gaze interaction. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, ETRA 2018, Warsaw, Poland, June 14-17, 2018*, Bonita Sharif and Krzysztof Krejtz (Eds.). Acm, 24:1–24:10. https://doi.org/10.1145/3204493.3204536

[3] Michael Barz and Daniel Sonntag. 2021. Automatic Visual Attention Detection for Mobile Eye Tracking Using Pre-Trained Computer Vision Models and Human Gaze. *Sensors* 21, 12 (Jan. 2021), 4143. https://doi.org/10.3390/s21124143 Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[4] Aljaž Božič, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. 2021. TransformerFusion: Monocular RGB Scene Reconstruction using Transformers. https://doi.org/10.48550/arXiv.2107.02191 arXiv:2107.02191 [cs].

[5] Ho Kei Cheng and Alexander G. Schwing. 2022. XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. http://arxiv.org/abs/2207.07115 arXiv:2207.07115 [cs].

[6] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. 2021. Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion. http://arxiv.org/abs/2103.07941 arXiv:2103.07941 [cs].

[7] Stijn De Beugher, Geert Brône, and Toon Goedemé. 2014. Automatic analysis of in-the-wild mobile eye-tracking experiments using object, face and person detection. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, Vol. 1. 625–633.

[8] Oliver Deane, Eszter Toth, and Sang-Hoon Yeo. 2022. Deep-SAGA: a deep-learning-based system for automatic gaze annotation from eye-tracking data. *Behavior Research Methods* (June 2022). https://doi.org/10.3758/s13428-022-01833-4

[9] Anna Gelencsér-Horváth, László Kopácsi, Viktor Varga, Dávid Keller, Árpád Dobolyi, Kristóf Karacs, and András Lőrincz. 2022. Tracking Highly Similar Rat Instances under Heavy Occlusions: An Unsupervised Deep Generative Pipeline. *Journal of Imaging* 8, 4 (April 2022), 109. https://doi.org/10.3390/jimaging8040109

[10] Benjamin Graham and David Novotny. 2020. RidgeSfM: Structure from Motion via Robust Pairwise Matching Under Depth Uncertainty. http://arxiv.org/abs/2011.10359 arXiv:2011.10359 [cs, eess] version: 1.

[11] Yuying Hao, Yi Liu, Yizhou Chen, Lin Han, Juncai Peng, Shiyu Tang, Guowei Chen, Zewu Wu, Zeyu Chen, and Baohua Lai. 2022. EISeg: An Efficient Interactive Segmentation Tool based on PaddlePaddle. http://arxiv.org/abs/2210.08788 arXiv:2210.08788 [cs].

[12] László Kopácsi, Árpád Dobolyi, Áron Fóthi, Dávid Keller, Viktor Varga, and András Lőrincz. 2021. RATS: Robust Automated Tracking and Segmentation of Similar Instances. In *Artificial Neural Networks and Machine Learning – ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part III*. Springer-Verlag, Berlin, Heidelberg, 507–518. https://doi.org/10.1007/978-3-030-86365-4_41

[13] Niharika Kumari, Verena Ruf, Sergey Mukhametov, Albrecht Schmidt, Jochen Kuhn, and Stefan Küchemann. 2021. Mobile Eye-Tracking Data Analysis Using Object Detection via YOLO v4. *Sensors* 21, 22 (2021). https://doi.org/10.3390/s21227668

[14] Kuno Kurzhals. 2021. Image-Based Projection Labeling for Mobile Eye Tracking. In *ACM Symposium on Eye Tracking Research and Applications*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3448017.3457382

[15] Kuno Kurzhals, Cyrill Fabian Bopp, Jochen Bässler, Felix Ebinger, and Daniel Weiskopf. 2014. Benchmark Data for Evaluating Visualization and Analysis Techniques for Eye Tracking for Video Stimuli. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (Beliv '14)*. Association for Computing Machinery, New York, NY, USA, 54–60. https://doi.org/10.1145/2669557.2669558 event-place: Paris, France.

[16] Eduardo Manuel Silva Machado, Ivan Carrillo, Miguel Collado, and Liming Chen. 2019. Visual Attention-Based Object Detection in Cluttered Environments. In *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. 133–139. https://doi.org/10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00064

[17] Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. 2014. Exploring a Model of Gaze for Grounding in Multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction (Icmi '14)*. Association for Computing Machinery, New York, NY, USA, 247–254. https://doi.org/10.1145/2663204.2663275 event-place: Istanbul, Turkey.

[18] Alexey Merzlyakov and Steve Macenski. 2021. A Comparison of Modern General-Purpose Visual SLAM Approaches. https://doi.org/10.48550/arXiv.2107.07589 arXiv:2107.07589 [cs].

[19] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. 2020. Atlas: End-to-End 3D Scene Reconstruction from Posed Images. https://doi.org/10.48550/arXiv.2003.10432 arXiv:2003.10432 [cs].

[20] Karen Panetta, Qianwen Wan, Aleksandra Kaszowska, Holly A. Taylor, and Sos Agaian. 2019. Software Architecture for Automating Cognitive Science Eye-Tracking Data Analysis and Object Annotation. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 268–277. https://doi.org/10.1109/thms.2019.2892919

[21] Thies Pfeiffer, Patrick Renner, and Nadine Pfeiffer-Leßmann. 2016. EyeSee3D 2.0: Model-Based Real-Time Analysis of Mobile Eye-Tracking in Static and Dynamic Three-Dimensional Scenes. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (Etra '16)*. Association for Computing Machinery, New York, NY, USA, 189–196. https://doi.org/10.1145/2857491.2857532 event-place: Charleston, South Carolina.

[22] Daniel F. Pontillo, Thomas B. Kinsman, and Jeff B. Pelz. 2010. SemantiCode: Using Content Similarity and Database-Driven Matching to Code Wearable Eyetracker Gaze Data. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (Etra '10)*. Association for Computing Machinery, New York, NY, USA, 267–270. https://doi.org/10.1145/1743666.1743729 event-place: Austin, Texas.

[23] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. 2021. Kimera: from SLAM to Spatial Perception with 3D Dynamic Scene Graphs. https://doi.org/10.48550/arXiv.2101.06894 arXiv:2101.06894 [cs].

[24] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. 2022. SimpleRecon: 3D Reconstruction Without 3D Convolutions. In *Computer Vision – ECCV 2022 (Lecture Notes in Computer Science)*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 1–19. https://doi.org/10.1007/978-3-031-19827-4_1

[25] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. 2020. f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation. http://arxiv.org/abs/2001.10331 arXiv:2001.10331 [cs].

[26] Ömer Sümer, Patricia Goldberg, Kathleen Stürmer, Tina Seidel, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2018. Teacher's Perception in the Classroom. *CoRR* abs/1805.08897. arXiv:1805.08897 http://arxiv.org/abs/1805.08897

[27] Zachary Teed and Jia Deng. 2020. DeepV2D: Video to Depth with Differentiable Structure from Motion. https://doi.org/10.48550/arXiv.1812.04605 arXiv:1812.04605 [cs].

[28] Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. 2012. Gaze Guided Object Recognition Using a Head-Mounted Eye Tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications (Etra '12)*. Association for Computing Machinery, New York, NY, USA, 91–98. https://doi.org/10.1145/2168556.2168570 event-place: Santa Barbara, California.

[29] Takumi Toyama and Daniel Sonntag. 2015. Towards Episodic Memory Support for Dementia Patients by Recognizing Objects, Faces and Text in Eye Gaze. In *KI 2015: Advances in Artificial Intelligence (Lecture Notes in Computer Science)*, Steffen Hölldobler, Rafael Peñaloza, and Sebastian Rudolph (Eds.). Springer International Publishing, Cham, 316–323. https://doi.org/10.1007/978-3-319-24489-1_29

[30] Karan Uppal, Jaeah Kim, and Shashank Singh. 2022. Decoding Attention from Gaze: A Benchmark Dataset and End-to-End Models. In *NeuRIPS 2022 Workshop on Gaze Meets ML*. https://openreview.net/forum?id=1Ty3Xd9HUQv

[31] Viktor Varga and András Lőrincz. 2021. Fast Interactive Video Object Segmentation with Graph Neural Networks. http://arxiv.org/abs/2103.03821 arXiv:2103.03821 [cs].

[32] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. 2021. Deep Two-View Structure-from-Motion Revisited. http://arxiv.org/abs/2104.00556 arXiv:2104.00556 [cs].

[33] Julian Wolf, Stephan Hess, David Bachmann, Quentin Lohmeyer, and Mirko Meboldt. 2018. Automating areas of interest analysis in mobile eye tracking experiments based on machine learning. *Journal of Eye Movement Research* 11, 6 (Dec. 2018). https://doi.org/10.16910/jemr.11.6.6 Section: Articles.

[34] Guangkai Xu, Wei Yin, Hao Chen, Chunhua Shen, Kai Cheng, Feng Wu, and Feng Zhao. 2022. Towards 3D Scene Reconstruction from Locally Scale-Aligned Monocular Video Depth. https://doi.org/10.48550/arXiv.2202.01470 arXiv:2202.01470 [cs].

[35] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. 2020. Learning to Recover 3D Scene Shape from a Single Image. https://doi.org/10.48550/arXiv.2012.09365 arXiv:2012.09365 [cs].

[36] L.H. Yu and M. Eizenman. 2004. A new methodology for determining point-of-gaze in head-mounted eye tracking systems. *IEEE Transactions on Biomedical Engineering* 51, 10 (Oct. 2004), 1765–1773. https://doi.org/10.1109/tbme.2004.831523

[37] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. 2022. NICE-SLAM: Neural Implicit Scalable Encoding for SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.