2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

ReViSE : Proposal of Framework which Enables Seamless and Flexible Integration of Real and Virtual Objects for Video See-Through MR

Koki Suzuki[1], Eito Yanagisawa[1], Hitoshi Iyatomi[2] and Sousuke Nakamura[2]

***Abstract*— Recent advancements in deep learning technology have significantly improved the performance of instance segmentation to a practical level. This technology enables the detection, segmentation, and extraction of object regions, offering substantial potential for applications in mixed reality (MR). While most research has focused on detection and segmentation, the application of extraction in realizing MR has received limited attention. In this paper, we propose a framework called ReViSE (Real and Virtual Seamless Editor), which integrates instance segmentation with virtual reality (VR) technology to deliver a wide range of MR experiences. This framework generates diverse MR visuals by applying instance segmentation on original images captured by a camera, and then replacing specified arbitrary object regions with virtual objects. Then, the MR visuals are presented through a head-mounted display (HMD) to provide users with a highly immersive visual experience. Evaluation experiments with a basic implementation show a processing time of 52.67 ms/frame and a display performance of 32.73 fps. The framework has also demonstrated its ability to accurately extract target objects and deliver a high-quality visual experience.**

I. INTRODUCTION

The rapid advancements in machine learning in recent years have led to the widespread use of image recognition technologies in various applications, such as autonomous driving and facial recognition. Instance segmentation is an advanced machine-learning technique that combines object detection and segmentation, enabling the precise identification of individual objects in an image at the pixel level, accurately segmenting contours, and assigning appropriate labels to the recognized objects [1]. Recent research has significantly improved both real-time performance and segmentation accuracy. YOLACT, proposed by Daniel et al. [2], performs parallel generation of non-local prototype mask sets across an entire image and predicts mask coefficients for each instance. Instance masks are generated by linearly combining these prototypes and mask coefficients. This model achieved 0.299 mask AP and 33.5 fps on the COCO dataset, demonstrating real-time instance segmentation. The Swin Transformer, proposed by Ze et al. [3], uses a hierarchical structure that progressively merges adjacent patches with deep transformer layers and bridges maps between attention layers. This approach achieved high-precision segmentation with 0.587 box AP and 0.511 mask AP on COCO datasets. These advancements have significantly increased the practicality of instance segmentation, leading to its application in various fields, including the diagnosis of tumors and abnormalities in the medical field [4], target detection in autonomous driving [5], and monitoring systems for construction site activities [6].

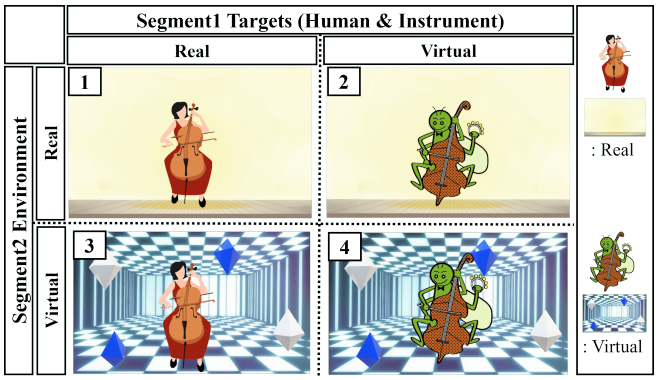

Fig. 1. Example of image representation using real and virtual objects.

We focused on the potential of object extraction using this technology and came up with its application in mixed reality (MR). MR is a technology that blends elements of virtual reality (VR) and the real world, offering a high level of immersion and versatility. Its application spans diverse fields, including healthcare [7], education [8], and training simulators [9]. However, there has been limited research on MR technology that incorporates object segmentation, and frameworks considering object selection are still scarce. Our concept is to selectively combine real and virtual objects, allowing for flexible configurations to create various expression patterns, as illustrated in Fig. 1, to provide MR images with superior expressiveness and versatility.

In this paper, we propose ReViSE (Real and Virtual Seamless Editor) to realize this concept. This is an innovative video see-through framework that generates diverse MR visuals by using instance segmentation on original images captured by a camera, replacing only specified arbitrary object regions (considering the background as one of the object regions) with virtual objects, and presenting them through a head-mounted display (HMD).

We developed a prototype system designed for live concerts and dance performances based on ReViSE, conducted evaluation experiments, and discussed its feasibility. The contributions of this paper are as follows:

- We propose ReViSE, a framework that generates diverse

[1]Koki Suzuki and Eito Yanagisawa are with the Graduate School of Science and Engineering, Hosei University as a master's student, Tokyo, Japan (e-mail:{koki.suzuki.9r, eito.yanagisawa.6t}@stu.hosei.ac.jp).

[2]Hitoshi Iyatomi and Sousuke Nakamura are with the Faculty of Science and Engineering, Hosei University as a professor, Tokyo, Japan (e-mail:{iyatomi, snakamura}@hosei.ac.jp).

MR visuals by replacing specified arbitrary object regions in the original image with virtual objects.
- The prototype based on ReViSE achieved 32.73 fps while performing moderate object extraction with selection, which has a potential to meet practical performance by further improvement in extraction accuracy and processing speed.

II. RELATED WORK

Yu et al. introduced a system employing chroma keying to extract object regions in the real world, enabling a real-time environment where real and virtual objects coexist [10]. However, this approach is limited to representing only patterns 3 and 4 in Fig. 1. Furthermore, due to restrictions on the types of objects that can be extracted and the unnatural appearance of the results, the method is not well-suited for practical on-site applications, such as live concerts. Michael et al. achieved MR visuals by using a depth sensor to extract the foreground [11]. While this method facilitates object region extraction in natural environments, it is restricted to expressing pattern 3 in Fig. 1. Additionally, it suffers from unclear boundaries due to camera and object positioning, making it challenging to accurately extract the foreground in dynamic scenes or those with multiple foreground elements. Pierre et al. developed a framework leveraging semantic segmentation in video streaming, allowing for the recognition and extraction of the user's body and other participants for integration into a virtual environment [12]. Similarly, Ester et al. conducted a comparative evaluation of foreground extraction using different segmentation techniques based on color and deep learning [13]. These studies demonstrated the effectiveness of deep learning-based segmentation methods for foreground extraction. However, they were limited to achieving pattern 3 in Fig. 1 and focused solely on human bodies, without addressing the generalizability to other types of objects. Dongdong et al. proposed a system using instance segmentation to extract objects from a desk in camera footage, selectively integrating them into a virtual environment [14].

This system provides the option to display recognized objects as either real or virtual, but it is capable of representing only patterns 3 and 4 in Fig. 1. Moreover, it targets objects near the HMD wearer and does not account for distant objects. In summary, the frameworks proposed in previous studies for achieving MR through object region extraction have not comprehensively considered all the expressive patterns illustrated in Fig. 1.

III. REVISE

The proposed ReViSE framework enables the replacement of arbitrary object regions, including the background, with virtual objects in images captured by a camera. Users or service providers can freely select which objects in the camera footage are to be replaced with virtual counterparts. By allowing flexible modification of the object regions to be replaced, the ReViSE framework facilitates a wide range of creative possibilities, enabling diverse MR expressions. For instance, it can create a scene where a single real object is replaced by a virtual one, resulting in a blend of real and virtual elements.

The configuration of the ReViSE framework for these functionalities is illustrated in Fig. 2. ReViSE comprises five core components: Image Acquisition, Instance Segmentation, Instance Selector, Hybrid Reality Synthesis, and Vision Sync Display. The process begins with Image Acquisition, where the original image is captured by a camera. In the Instance Segmentation stage, the image is analyzed to recognize, segment, and label objects. The Instance Selector then allows users or service providers to choose the labeled objects to be replaced with virtual counterparts. In the Hybrid Reality Synthesis step, these selected objects are substituted with virtual objects or backgrounds, producing a composite image that integrates both real and virtual elements. Finally, the Vision Sync Display presents the synthesized visuals to the user through a HMD. Due to the significant computational resources required for Instance Segmentation, Instance Selector, and Hybrid Reality Synthesis, the framework is designed to utilize a high-performance remote computer. Moreover, the HMD used for displaying visuals to the user is compatible with various devices, such as personal computers or smartphones, offering flexible operation and versatility.

IV. EVALUATION EXPERIMENT

In this section, we present a prototype of the ReViSE framework designed to provide a hybrid view of real objects and virtual environments, as illustrated in Pattern 3 of Fig. 1, specifically targeting human subjects at an on-site live concert. We tested this prototype in a live concert setting to evaluate its segmentation performance and processing speed through both quantitative and qualitative analyses.

A. Prototype Configuration

The prototype setup utilized a Stereolabs ZED mini camera and a Meta Quest2 headset as the display device. A remote computer, responsible for instance segmentation and other processing tasks, was equipped with a Core i7-13700KF processor, a GeForce RTX 4070 GPU, and 64GB of RAM. Each device was connected via USB cables. The ZED mini camera was mounted on the Meta Quest2 to capture images from the viewer's perspective, as shown in Fig. 3.

For instance segmentation, we prioritized real-time performance by employing RTMDet, a state-of-the-art model developed. [15]. This model utilizes large kernel depthwise convolutions and dynamic label assignment with soft labels, achieving a performance of over 300 fps and 0.528 in AP on the MSCOCO dataset when running on an NVIDIA 3090 GPU. To further enhance real-time performance in our prototype, we chose the RTMDet-Ins-tiny model, the fastest inference model within the RTMDet family.

B. Processing Details of the Prototype

Figure 4 shows the details of the ReViSE prototype. The Instance Segmentation process analyzes each frame of

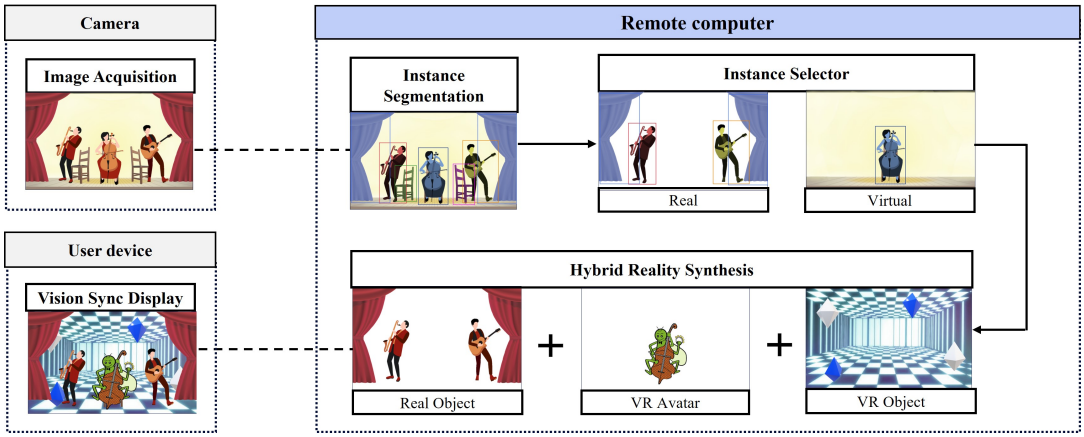

Fig. 2. Configuration diagram of ReViSE.

Fig. 3. Meta Quest2 with ZED mini camera.

the image captured by the camera to extract objects. The Instance Selector then refers to a predefined list of object labels designated for replacement with virtual objects. It generates images by retaining only the objects whose class labels match those identified by the Instance Segmentation. These processes are executed on the GPU. Hybrid Reality Synthesis is implemented using the Unity game engine [16]. Two screens are configured for the left and right eyes, with the HMD's viewpoint camera positioned directly in front of them. The screens are curved to align with the viewer's perspective through the HMD, ensuring an equal distance from the viewpoint to all parts of the screen, as illustrated in Fig. 5. The extracted images are projected onto these screens, and by combining them with virtual objects (in this case, a virtual background), the MR visuals are generated.

Furthermore, as shown in Fig. 4, the process from Image Acquisition to Image Sender and the subsequent process are executed separately, and the data communication between Image Sender and Image Receiver is asynchronous using UDP communication. Therefore, the next Image Acquisition is performed immediately after Image Sender.

C. Experimental Environment and Criteria

The evaluation experiments were conducted in a simulated front-row live concert environment, where one or two individuals were designated as extraction targets, positioned at distances of 1.5m and 3.0m from the camera. The input image resolution was set to 1280 × 640 pixels, combining the left and right images. For the virtual background, we used the "Lowpoly Environment - Nature Free - MEDIEVAL FANTASY SERIES" asset[1] from the Unity Asset Store, which was composited with the extracted individuals. In this prototype, data communication between devices (camera, computing resources, and HMD) was conducted through wired connections. The visuals presented to the user were evaluated in terms of segmentation performance and processing speed.

1) Segmentation Performance: For the quantitative evaluation of Instance Segmentation, we used the results of Segment Anything (SAM) [17], a state-of-the-art high-precision segmentation method, as a silver standard, serving as an approximate gold standard.

SAM is a method reported to have high object extraction accuracy, with an average MaskAP of 0.465 and a maximum MaskAP of 0.617 on the COCO dataset, and it also achieves exceptionally high detection accuracy for people and other objects in our task. However, SAM requires that the bounding boxes surrounding the objects or representative points within the objects be provided manually or generated by another machine learning model in advance. Additionally, in our experimental environment, the time required for segmentation by SAM (an average of 34.36 ms per frame) is much longer than that of the RTMDet-Ins-tiny model (7.42 ms per frame) used in the prototype. As a result, SAM currently does not meet our requirements for automation and computational efficiency. On the other hand, SAM's high-performance instance segmentation results can serve as a useful evaluation metric. Given the prohibitive cost of manually labeling a large number of image frames to evaluate the prototype's performance, SAM was adopted as the silver standard.

To evaluate extraction accuracy, we employed several metrics: Precision, Recall, F1 score, and Boundary IoU (BIoU) [18]. Precision measures the proportion of correctly predicted positive object regions, while Recall measures the proportion of true positive object regions relative to the ground truth labels. The F1 score provides a balanced evaluation by calculating the harmonic mean of Precision and Recall. BIoU

[1]https://assetstore.unity.com/packages/3d/environments/lowpoly-environment-nature-free-medieval-fantasy-series-187052.

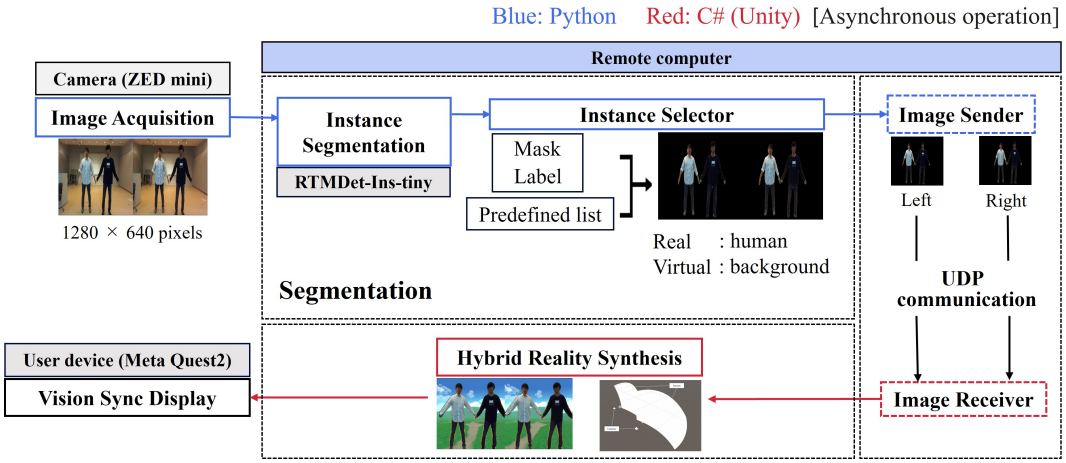

Fig. 4. Details of the ReViSE prototype developed in this experiment.

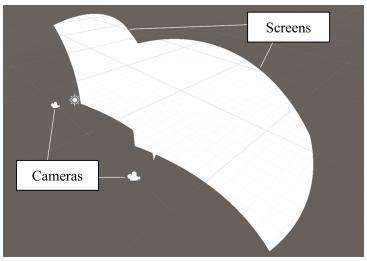

Fig. 5. Two screens and two cameras placed in the Unity game engine.

specifically evaluates the accuracy of the object's shape near its boundaries. It includes a distance parameter d , which adjusts the sensitivity of BIoU to boundary pixels; a smaller d value places more emphasis on the boundary pixels. In our evaluation, we set d=15 pixels. Let P denote the predicted mask and G the ground truth mask. P_d and G_d represent the mask pixels within a distance d from their respective boundaries. BIoU is defined by the following equation:

$$\mathrm{BIoU} = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|} \quad (1)$$

2) Processing Speed (Video Delay and Smoothness): To evaluate the processing speed, we measured the time required for each frame to complete Image Acquisition, Segmentation[2], UDP communication (i.e., communication time between components), and Hybrid Reality Synthesis in scenarios with one and two targets at varying distances. By comparing these times against predefined evaluation criteria, we assessed the image delay between the visual output shown to the user and the accompanying live sound. For this evaluation, we used an acceptable delay range of 20 to 30 ms, as established by Schuett's study [19], which investigated the perceptual differences between visual and auditory stimuli in human perception.

[2]Instance Segmentation and Instance Selector were combined into one measurement with Segmentation because instance selection and mask drawing are performed at the same time.

Fig. 6. Input image acquired from camera.

Additionally, we calculated the FPS from the interval time of Image Acquisition because the smoothness of the image is also an important factor related to the user's video experience. In Fig. 4, the upper process (Image Acquisition, Instance Segmentation, Instance Selector, Image Sender) and the lower process (Image Receiver, Hybrid Reality Synthesis, Vision Sync Display) is conducted asynchronously. Therefore, this interval time corresponds to the upper process, from which the FPS was calculated. Here, it should be noted that the processing time for both the upper and lower processes did not have any variation, which means that this FPS certainly represents the smoothness of the image.

For these measurements, we considered the latency of data communication between the wired devices to be negligible (effectively zero).

V. RESULT

A. Segmentation performance

Figure 6 shows the input images captured by the camera, while Fig. 7 presents the results with and without the Instance Selector's selection feature at distances of 1.5m and 3.0m. Video examples are available in the following GitHub repository: https://github.com/MRliveN417/MRlive.git.

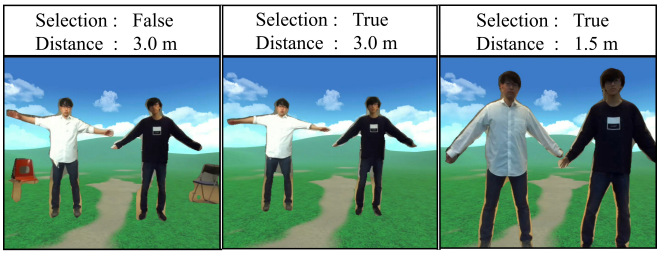

Fig. 7. Output results (Left: Instance Selection is false, distance 3.0m Center: Instance Selection is true, distance 3.0m Left: Instance Selection is true, distance 1.5m).

In Fig. 7, when Instance Selection is enabled, only people are selected and extracted from the detected objects (people and chairs), confirming that the prototype functions as intended. At a distance of 3.0m, the overall extraction accuracy was low, particularly around the face, shoulders, and lower body, whereas at a distance of 1.5m, the extraction was relatively accurate.

Figure 8 compares the extraction results of the prototype with those obtained by SAM, which was used as the silver (semi-gold) standard. Table I provides the Precision, Recall, F1 score, and BIoU values calculated from 150 frames.

The experiment revealed that Precision and BIoU decreased as the distance to the subject increased, indicating an expansion of the detected region and a resulting decline in extraction accuracy, highlighting areas for further improvement. Nevertheless, the high extraction accuracy achieved by SAM, used as the silver standard, suggests that these challenges may be mitigated in the near future with advancements in computational power.

B. Processing Speed (Video Delay and Smoothness)

Table II summarizes each processing time per frame for each condition, the total time, and the interval time of Image Acquisition until the next frame is acquired. The values are averages over 100 frames processed.

Table II demonstrates the system's stable operation speed, with no significant differences in processing time across different conditions. This reliability ensures consistent performance under various circumstances. The fastest overall processing speed was 52.67 ms/frame. Compared to the benchmarks established by Schuett's research, even the shortest total processing time takes 1.8 to 2.6 times longer, indicating noticeable video latency may cause user inconvenience at this stage. UDP communication accounted for approximately 44 to 51% of the total processing time, indicating it as the rate-limiting step. The interval time of Image Acquisition achieved a maximum of 30.55 ms/frame, corresponding to 32.73 fps. This value exceeds the commonly used standard of 30 fps, such as in Zoom, indicating that the system can provide sufficiently smooth video.

VI. DISCUSSION

The developed prototype displays arbitrarily selected and extracted objects from input images within a virtual environment, which has potential to provide users with an immersive live experience. Moreover, the prototype offers a wide range of expression patterns, as shown in Fig. 1, which can be realized by simply changing the selected objects and the display format, inspiring new possibilities in the field of mixed reality. However, the prototype still needs to improve, such as low extraction accuracy for distant objects and complex, small shapes such as facial features. These issues could be improved by utilizing telephoto cameras or super-resolution techniques to enlarge the size of the target objects. Additionally, applying more accurate segmentation methods could further enhance performance.

The evaluation of processing speed against the benchmark revealed that the prototype experienced unacceptable video latency, highlighting the need for faster processing. The UDP communication, identified as the rate-limiting step, could be accelerated by multiprocessing or multithreading and

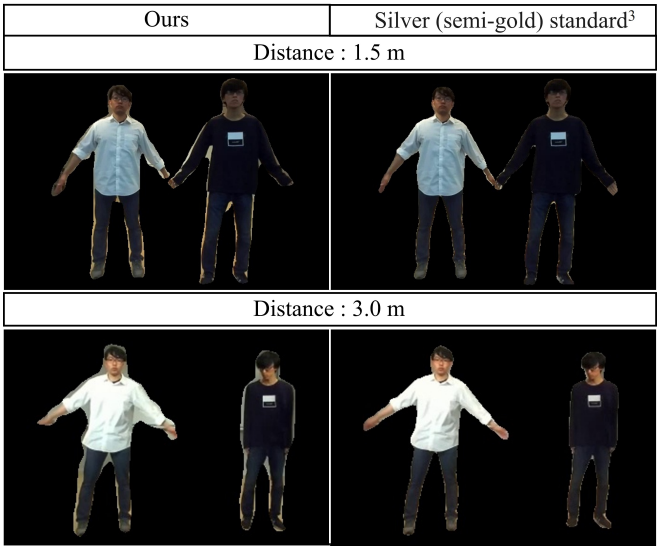

Fig. 8. Instance segmentation results for persons at different distances.

TABLE I
SUMMARY OF DETECTION PERFORMANCE.

Distance	Precision	Recall	F1	BIoU
1.5m	0.917 ± 0.006	0.976 ± 0.004	0.946 ± 0.003	0.813 ± 0.012
3.0m	0.853 ± 0.007	0.966 ± 0.008	0.906 ± 0.005	0.769 ± 0.012

TABLE II
TIME MEASUREMENTS IN THE PROTOTYPE.

Number of Targets		One person		Two persons	
Distance	(m)	1.5	3.0	1.5	3.0
Image Acquisition		7.97	9.57	8.82	8.45
Segmentation		16.11	16.13	14.95	16.53
UDP communication	(ms/frame)	27.78	23.05	28.80	25.06
Hybrid Reality Synthesis		4.22	3.92	4.28	4.03
Total		56.08	52.67	56.85	54.07
Interval time of Image Acquisition	(ms/frame)	30.82	30.55	31.20	30.68
FPS	(frame/s)	32.45	32.73	32.05	32.59

[3]Segmentation results using SAM [17]. This image was created step by step by giving us the bounding boxes which is the object area in advance.

reducing the amount of transmitted data through image compression. The segmentation process could be significantly accelerated by upgrading the current mid-range GPUs to more advanced ones. This upgrade could have a substantial impact on the overall performance of the system. On the other hand, the output video achieved 32.73 fps, ensuring a practical level of smoothness. This indicates the system can provide adequate video quality even for dynamic activities such as dance performances.

VII. LIMITATION

The evaluated prototype used wired connections between devices because there is a drawback when processing tasks such as instance segmentation on wireless connections (server) where latency is possible. Additionally, the Instance Selector currently only allows for selecting objects by category without considering the selection of individual objects within the same category. The evaluation experiments were conducted under ideal conditions, without any obstacles obstructing the target. However, it's important to note that in real-world scenarios, target occlusion could occur, potentially leading to incomplete object extraction.

VIII. CONCLUSION

In this paper, we proposed ReViSE, a new video see-through framework that enables diverse visual representations by integrating instance segmentation with MR technology. We developed and evaluated a prototype as a proof of concept. The results demonstrated that it is possible to select and display arbitrary objects in a virtual space, allowing for highly flexible visual representation. The prototype also achieved high-speed performance, ensuring a smooth MR experience. However, challenges remain concerning video latency and the accuracy of object extraction for distant or complex-shaped objects. We plan to address these issues in future work.

ACKNOWLEDGMENT

We would like to thank T. Ito for the useful discussions. We also thank A. Prajina for discussions and carefully proofreading the manuscript.

REFERENCES

[1] R. Sharma, M. Saqib, C. T. Lin, and M. Blumenstein, A Survey on Object Instance Segmentation, SN COMPUT. SCI. , vol. 3, 2022, pp. 499, doi: 10.1007/s42979-022-01407-3.

[2] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, YOLACT: Real-time Instance Segmentation, Proceeding of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 9157-9166.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012-10022.

[4] I. Tanasković, S. Ičagić, I. Šolić, and B. Rakić., Adapting YOLOv8 for Kidney Tumor Segmentation in Computed Tomography, 2024 9th International Conference on Smart and Sustainable Technologies (SpliTech), 2024, doi: 10.23919/SpliTech61897.2024.10612634.

[5] L. Guan and X. Yuan, Instance Segmentation Model Evaluation and Rapid Deployment for Autonomous Driving Using Domain Differences, in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 4, pp. 4050-4059, April 2023, doi: 10.1109/TITS.2023.3236626.

[6] R. Bai, M. Wang, Z. Zhang, J. Lu and F. Shen, Automated Construction Site Monitoring Based on Improved YOLOv8-seg Instance Segmentation Algorithm, in IEEE Access, vol. 11, pp. 139082-139096, 2023, doi: 10.1109/ACCESS.2023.3340895.

[7] L. Chen, T. W. Day, W. Tang and N. W. John, Recent Developments and Future Challenges in Medical Mixed Reality, 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Nantes, France, 2017, pp. 123-135, doi: 10.1109/ISMAR.2017.29.

[8] X. Xu, A. Puggioni, D. Kilroy and A. G. Campbell, User Experience of Collaborative Co-located Mixed Reality: a User Study in Teaching Veterinary Radiation Safety Rules, 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, 2023, pp. 583-590, doi: 10.1109/ISMAR59233.2023.00073.

[9] T. Laudien, J. M. Ernst and B. Isabella Schuchardt, Implementing a Customizable Air Taxi Simulator with a Video-See-Through Head-Mounted Display – A Comparison of Different Mixed reality Approaches, 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), Portsmouth, VA, USA, 2022, pp. 1-10, doi: 10.1109/DASC55683.2022.9925870.

[10] Y. Zhu, K. Zhu, Q. Fu, X. Chen, H. Gong, and J. Yu, SAVE: Shared augmented virtual environment for real-time mixed reality applications, Proceedings of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry, VRCAI'16, pp. 13–21, 2016, doi: 10.1145/3013971.3013979.

[11] M. Rauter, C. Abseher and M. Safar, Augmenting Virtual Reality with Near Real World Objects, 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 2019, pp. 1134-1135, doi: 10.1109/VR.2019.8797873.

[12] P. -O. Pigny and L. Dominjon, Using CNNs For Users Segmentation In Video See-Through Augmented Virtuality, 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), San Diego, CA, USA, 2019, pp. 229-2295, doi: 10.1109/AIVR46125.2019.00048.

[13] E. Gonzalez-Sosa, P. Pérez, R. Tolosana, R. Kachach and A. Villegas, Enhanced Self-Perception in Mixed Reality: Egocentric Arm Segmentation and Database With Automatic Labeling, in IEEE Access, vol. 8, pp. 146887-146900, 2020, doi: 10.1109/ACCESS.2020.3013016.

[14] D. Weng, W. He, S. Guo and D. Li, Functional-Penetrated Interactive System Towards Virtual-Real Fusion Environments, 2022 14th International Conference on Signal Processing Systems (ICSPS), Jiangsu, China, 2022, pp. 842-850, doi: 10.1109/ICSPS58776.2022.00151.

[15] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, RTMDet: An Empirical Study of Designing Real-Time Object Detectors, arXiv preprint arXiv:2212.07784, 2022.

[16] Unity Technologies. unity, https://unity.com/ja.

[17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár and R. Girshick, Segment Anything, arXiv preprint arXiv:2304.02643, 2023.

[18] B. Cheng, R. Girshick, P. Dollar, A. C. Berg, and A. Kirillov, Boundary IoU: Improving Object-Centric Image Segmentation Evaluation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15334-15342

[19] N. Schuett, The Effects of Latency on Ensemble Performance, Bachelor Thesis, Dept. Center Comput. Res. Music Acoust., Stanford Univ., Stanford, CA, USA, 2002.