VirtualClassroom: A Lecturer-Centered Consumer-Grade Immersive Teaching System in Cyber–Physical–Social Space

Tianyu Shen, Shi-Sheng Huang, Deqi Li, Zhiyuan Lu, Fei-Yue Wang, *Fellow, IEEE*, and Hua Huang, *Senior Member, IEEE*

***Abstract*—Lecturers, as the guidance of the classroom, play a significant role in the teaching process. However, the lecturers' sense of space immersion has been ignored in current virtual teaching systems. In this article, we explore the cyber–physical–social intelligence for Edu-Metaverse in cyber–physical–social space and specially design a lecturer-centered immersive teaching system, taking the social and lecturers' factors into consideration. We call this system VirtualClassroom (V-Classroom). Specifically, we first introduce the cyber–physical–social system (CPSS) paradigm of V-Classroom so that the workflow is standardized and significantly simplified, and the systems can be constructed with off-the-shelf hardware. The key component of V-Classroom is a cyber-world representation of a physical-world classroom instrumented with sparse consumer-grade RGBD cameras for capturing the 3-D geometry and texture of the classrooms. We provide each V-Classroom lecturer with a physical device for sending 6DoF view-change messages and showing view-dependent content of the remote classroom. Following the above paradigm, we develop the V-Classroom algorithms, including V-Classroom depth algorithm (V-DA) and V-Classroom view algorithm (V-VA), to achieve the real-time rendering of remote classrooms. V-DA is dedicated to recovering accurate depth information of the classrooms while V-VA is devoted to real-time novel view synthesis. Finally, we illustrate our implemented CPSS-driven V-Classroom prototype, based on real-world classroom scenarios we collected, and discuss the main challenges and future direction.**

***Index Terms*—6DoF video, cyber–physical–social systems (CPSSs), educational metaverse, immersive teaching.**

I. Introduction

THE PROBLEM of time–space separation between lecturers and learners in remote teaching has been widely concerned by researchers [1], [2], [3]. Although recent online synchronous teaching technologies have effectively made up for the time separation problem by means of live video, the space separation still remains a challenge.

The Metaverse [4], [5], an emerging conception based on 5G network, virtual reality (VR) and other information technologies [6], [7], [8], [9], describes physical worlds and virtual worlds. The virtual worlds not only reflect exactly the physical worlds but also able to expand infinitely to form a superlarge space where the physical worlds and virtual worlds interact with each other [10]. From an engineering perspective, the Metaverse can be regarded as a specific realization of CPSS, which specially refers to three spaces (physical, cyber, and social spaces) and two worlds (physical and virtual worlds). The educational Metaverse (Edu-Metaverse) has the ability to provide an immersive teaching field and transcend the barriers of space separation [11], [12]. The visual immersion of lecturers and learners is regarded as the core consideration for the exploration of Edu-Metaverse. Nowadays, only a few VR-based immersive teaching systems have emerged, but they only take the immersion of learners into consideration and are characterized by specialized device customization, complex scene construction, and limited service life.

To meet the above challenges, this article explores the cyber–physical–social intelligence for Edu-Metaverse in cyber–physical–social spaces. A layered architecture of CPSS for Edu-Metaverse is illustrated in Fig. 1. Based on this architecture, a lecturer-centered consumer-grade immersive teaching system named VirtualClassroom (V-Classroom) is designed in a CPSS paradigm, taking the lecturers' factors (belonging sense to the classroom, teaching motivation, enthusiasm, etc.) and social factors (affordability, reproducibility, flexibility, etc., for education equity) into consideration. Meanwhile, V-Classroom also serves as a supporting part, which focuses on immersion, for Edu-Metaverse. In contrast to existing immersive teaching systems, the special advantages of V-Classroom mainly lie in two aspects.

1) V-Classroom is designed to be lecturer-centered, as a real-time 6DoF video communication system specially for the lecturers in remote teaching. Actually, the lecturers are the guidance of the classroom and key to the teaching quality [13]. Improving lecturers' sense of space immersion will help stimulate their teaching initiative and enthusiasm.
2) V-Classroom workflow is standardized and significantly simplified by means of a cyber–physical–social

Manuscript received 15 November 2022; accepted 29 November 2022. Date of publication 20 December 2022; date of current version 18 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61533019. This article was recommended by Associate Editor Y. Tang. *(Corresponding author: Hua Huang.)*

Tianyu Shen, Shi-Sheng Huang, Deqi Li, Zhiyuan Lu, and Hua Huang are with the School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China (e-mail: tianyu.shen@bnu.edu.cn; huangss@bnu.edu.cn; dqli@mail.bnu.edu.cn; zylu@mail.bnu.edu.cn; huahuang@bnu.edu.cn).

Fei-Yue Wang is with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feiyue.wang@ia.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2022.3228270.

Digital Object Identifier 10.1109/TSMC.2022.3228270

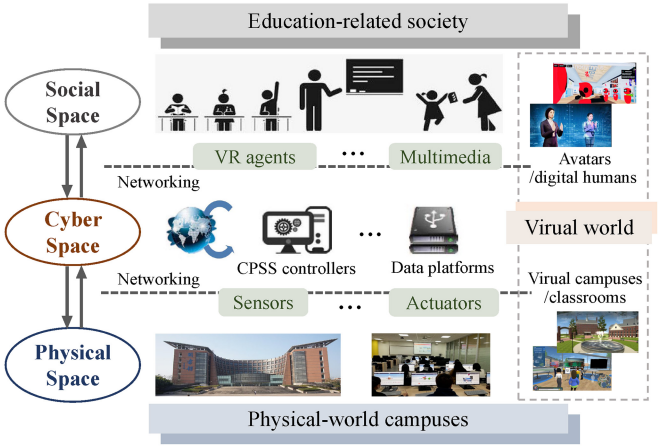


Fig. 1. Layered architecture of CPSS for Edu-Metaverse in cyber–physical–social space.

system (CPSS) paradigm. Through massive computational experiments in the cyber world, the interactions between the cyber world and physical world, and the consideration of social factors in the mental world, CPSS can achieve intelligent management and control of V-Classrooms, thus, reducing costs, improving reproducibility, and achieving off-the-shelf and flexible system construction.

To be specific, the key ingredient of V-Classroom is an abstract representation in the cyberspace of a classroom in physical space and it is used as the basic component of V-Classroom system. Sparse consumer-grade RGBD cameras are configured in the physical-world classroom for capturing the 3-D geometry and texture of the classroom scene and learner characters. And we provide each V-Classroom lecturer with a physical device for sending 6DoF view-change messages and showing view-dependent scene-characters content in real time. To achieve the real-time rendering of remote classrooms with learners, we further develop the V-Classroom algorithms including V-Classroom depth algorithm (V-DA) and V-Classroom view algorithm (V-VA). The V-DA is dedicated to recovering accurate depth information of the classrooms while V-VA is devoted to real-time novel view synthesis. Also, we provide a preliminary implementation of V-Classroom based on the real-world classroom scenarios we collected. In summary, this article has made three contributions as follows.

1) We propose the V-Classroom, a real-time 6DoF video communication system for the lecturers in remote teaching, following the CPSS paradigm. The V-Classroom is designed to be lecturer-centered and consumer-grade with a standardized workflow defined in the CPS space.
2) We develop the V-Classroom algorithms for achieving real-time rendering of remote classrooms with learner characters. The V-Classroom algorithms consist of a V-DA for acquiring accurate scene depth and a V-VA for synthesizing novel view frame.
3) Furthermore, we collect two certain real-world classroom scenarios and provide a preliminary V-Classroom implementation. The V-Classroom correctly preserves the scene changes of the classroom along with remote learners, allowing lecturers to perceive the learners' states and enhancing lecturers' sense of belonging to the teaching process.

The remainder of this article is organized as follows. In Section II, we summarize the related work of this study. In Section III, we introduce a CPSS architecture for Edu-Metaverse and provide a systematic description of V-Classroom in a CPSS paradigm. In Section IV, several challenges of V-Classroom are analyzed and our proposed V-Classroom algorithms are explained. In Section V, we provide the experiments and results of a preliminary V-Classroom implementation. Finally, Section VI concludes this article.

II. Related Work

In this section, we first introduce the related concepts and applications of CPSS. Then, we overview the related work on Edu-Metaverse and immersive teaching as well as illustrate the novelty of V-Classroom. Additionally, we survey the related research on indoor depth estimation and novel view syntheis according to the fields involved in V-Classroom algorithms.

A. Cyber–Physical–Social Systems

CPSS proposed by Wang [14] is defined as an extension of cyber–physical system (CPS) [15] with an incorporation of social factors such as human performance. The CPS describes a computation–communication–control integrated system that tightly conjoins and coordinates cyber and physical attributes [16], [17]. The design paradigms of CPS refer to a broad range of network-connected and physically aware systems embedding intelligent technologies in the cyber world into the physical world with computational nodes [18], [19].

In contrast to CPS, CPSS takes social factors, such as human performance [20], [21], into consideration and integrates cyber space, physical space, as well as social space. Actually, various terminologies and conceptualizations have emerged to represent the incorporation of social and human factors into CPS, such as cyber–physical–human systems (CPHSs) [22], [23], cyber–physical–social–thinking (CPST) hyperspace [24], [25], social–cyber–physical systems (SCPSs) [26], [27], [28], and so on. Nevertheless, the term CPSS has been conceived and adopted in most researches on the integration of CPS and social aspects. However, the ways of definition are not consistent because the usage of CPSS is dependent on the application fields.

CPSS results in a paradigm shift of intelligent complex systems and human societies by integrating the cyber space, physical space, and social space seamlessly. For guiding the corresponding physical systems and integrating multifaceted resources [29], CPSS has been effectively applied in many fields, including intelligent manufacturing [30], [31], energy [32] and power grid [33], [34], [35], intelligent transportation [36], smart vehicles [37], [38], enterprise management [14], [39], military operation [40], smart cities [41], [42], [43], et al.

In this article, we explore the cyber–physical–social intelligence for Edu-Metaverse in cyber–physical–social spaces

and specially design a lecturer-centered immersive teaching system in a CPSS paradigm, taking the lecturers' and social factors into consideration. The details on CPSS paradigm of V-Classroom are described in Section III.

B. Edu-Metaverse and Immersive Teaching

The Edu-Metaverse should be deployed to meet at least three characteristics which are high immersion, social interaction, and diversity.

1) *Immersion:* The Edu-Metaverse conceives a virtual world similar to the physical world by simulating the physical laws. The highly authentic virtual education world will enhance users' sense of belonging to the education process and enable immersive teaching [44].
2) *Interactivity:* Lecturers, learners, teaching resources, and learning environments are the basic elements of an educational scenario [45]. The interactivity among them is important for expanding the learning space, creating an almost realistic social space, and forming a sense of community.
3) *Diversity:* The rules in Edu-Metaverse should be free, open and flexible, unlike commercial games [46]. Lecturers and learners are allowed to create and communicate freely so as to form an infinite and diverse range of educational activities.

Recently, some Edu-Metaverse platforms have emerged. Immersive Journalism [47] provides the sensation of being present in the place by representing events on a spherical stage generated from real images that the user can control, so as to develop some collaboration activities for nurturing speaking skills. VoRtex [48] is primarily designed to support collaborative learning activities with the virtual environments and to support educational standards. VR-making and metaverse-linking for instructional content [49] are designed for preservice English teachers in instructional VR content design of K–12 and represent an open-source accessible solution developed using modern technology stack and metaverse concepts. Virtual worlds types for creating gameful experiences [50] are introduced to access the Metaverse for equal interact and educational opportunities. AViLab gamified system [51] is developed as an educational tool dedicated to experimentation and demonstration regarding an agent's features and basic principles. However, the current technologies are not mature enough to create an ideal Edu-Metaverse that completely meets all the required characteristics.

As for remote teaching, the space separation between lecturers and learners has been one of the most concerned and challenging problems [1], [2], [3]. And the Edu-Metaverse has shaped a visually immersive space field for remote courses [11], [12], [52]. Nowadays, few VR-based immersive teaching systems have emerged, but they only take learners' immersion into consideration and are characterized by specialized device customization, complex scene construction, and high costs. For example, Saïd Business School has constructed the first virtual meeting space of U.K., named the Oxford Hub for International Virtual Education (HIVE). This immersive classroom, centered around a high-definition video wall, blends the virtual reach with real engagement and employs cutting-edge technologies [53]. However, the lecturers' immersion is not enough, and such classrooms relying on exquisite equipment are costly to reproduce.

In this article, we propose a V-Classroom fully different from the existing immersive teaching systems. On the one hand, V-Classroom is designed to be a real-time 6DoF video communication system specially for the lecturers to improve lecturers' sense of space immersion in remote teaching. On the other hand, V-Classroom workflow is standardized and significantly simplified in the form of CPSS paradigm, which enables reducing costs, improving reproducibility, and achieving off-the-shelf and flexible system construction.

C. Novel View Synthesis

In recent years, novel view synthesis has always been an important concern in the field of both computer graphics (CG) and computer vision (CV). The related technologies mainly consist of model-based rendering (MBR) and image-based rendering (IBR).

The early MBR methods concentrate on applying the CG technologies to realize geometric modeling and graphic rendering [54], [55]. Such methods are only applicable to simple scenes due to the high complexity and high requirements for hardware devices. Recent MBR methods are committed to achieve novel view synthesis by exploiting CV approaches to build explicit 3-D geometric representations, such as voxel mesh [56], octree [57], point cloud [58], triangle mesh [59], and so on, from single or multiple images. However, they are still computationally intensive due to the explicit geometric inference and are prone to the loss of partial information during geometric estimation. Moreover, most of the learning-based MBR methods additionally require the 3-D geometry ground truth to train deep networks and cannot be generalized to unseen scenes [60], [61].

IBR methods [62] explicitly or implicitly encode the scenes based on the single or multiview images, and then render novel-view images from the explicit or implicit 3-D representations. Such methods are more suitable for real-time view synthesis of dynamic scenes because their computational complexity is not affected by the scenes and the authenticity is stronger. IBR methods can be divided into two categories according to whether they depend on geometric priors or not. The IBR methods that do not rely on geometric priors mainly refer to light field rendering [63], [64], [65]. However, the light field methods require a collection of extremely dense reference views, which usually relies on professional light field cameras or camera arrays. The IBR methods relying on geometric prior concentrate on achieving novel view synthesis based on the explicit or implicit geometric representation. The IBR with explicit geometric representation are similar to the CV-based MBR methods. The implicit geometric representation mainly refers to the depth information of the scenes. The depth-based IBR (DIBR) methods are able to synthesize high-quality novel views with the requirement of only a few reference views with depth maps. DIBR methods achieve a tradeoff between computational complexity and synthesis quality, resulting in a

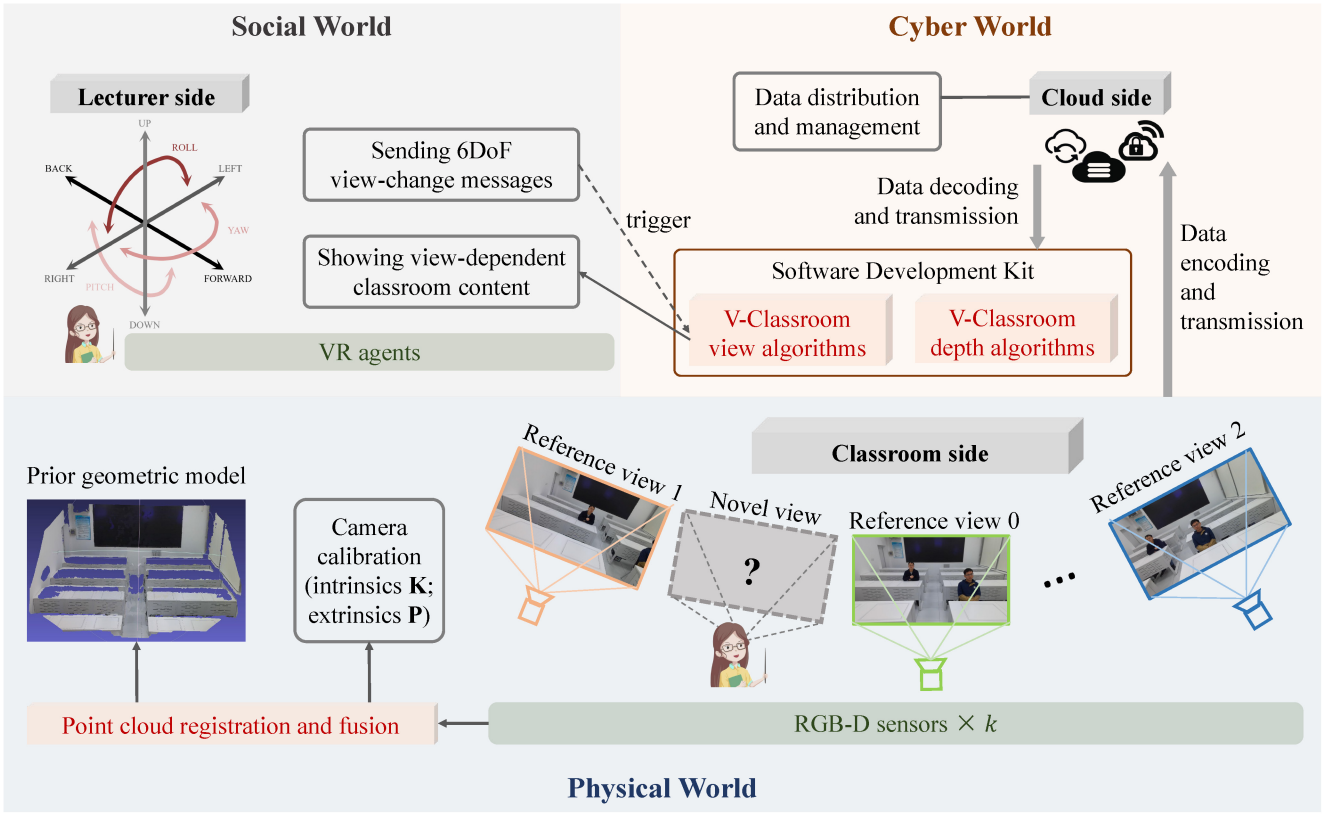


Fig. 2. Data and workflow of the V-Classroom system deployed in a CPSS paradigm, which encompasses a classroom side in the physical world, a lecturer side in the social world, and a cloud side in the cyber world.

wider range of applications. Recently, deep learning has been introduced to replace the manually designed phases of DIBR pipelines [66]. However, the view synthesis quality of DIBR is still extremely sensitive to the accuracy of depth information.

In this article, we incorporate the DIBR methods for view synthesis in V-Classroom. However, it is challenging to obtain high-quality depth maps for real-world classroom scenes with frequent textureless and texture-repeated areas. Thus, we explore the MBR methods for a prior geometric modeling of classroom scenes as well as indoor depth estimation and completion approaches, so as to realize more reliable depth guidance for view synthesis.

D. Indoor Depth Estimation

The challenges of indoor depth estimation are from the frequent textureless surfaces, such as large-area walls and floors, and various objects that are arbitrarily arranged in the near field. Also, the indoor depth tends to distribute unevenly in either near or far ranges (e.g., the zoomed-in views of desks or ceilings) while the depth distribution of outdoor scenes tends to be more uniform across near to far ranges on roads. To meet above challenges, the mainstream approaches for indoor depth estimation can be classified into active light sensing technologies and passive depth estimation methods.

Active light sensing technologies are dedicated to acquiring depth, relying on the auxiliary optical signal actively projected by the sensors [67]. The most common sensors include time of flight (ToF) sensors [68], [69] and structured light sensors [70], [71]. The light source used in active light vision has a fixed structure and optical properties, and it does not depend on the feature matching between color images. These lead to a good perception ability for textureless surfaces and a high acquisition efficiency. Active light sensing technologies greatly improve the accuracy of depth estimation for indoor scenes with massive textureless surfaces. However, the depth perception performance of active light sensors is easily affected by illumination, black surface, transparent materials, and other challenging factors.

Passive depth estimation methods are generally divided into MVS-based depth estimation and image-based depth regression. On the one hand, MVS-based depth estimation methods are realized by a series of stages, including feature extraction, feature matching, matching cost calculation, cost aggregation, depth estimation, and depth refinement, from input multiview or frame sequential images. Global stereo matching methods utilize graph cut algorithms [72] and dynamic programming [73] and so on to solve the feature mismatching problem of textureless regions, while local stereo matching methods solve the problem by means of feature operator optimization [74], segmentation-based region matching [75], phase matching [76], and so on. But most of these methods are inefficient, cumbersome, and unable to obtain dense depth maps. On the other hand, image-based depth regression methods make a breakthrough in depth estimation performance by virtue of CNN-based abstract feature extraction. No longer do

such methods use feature matching between images, so as to overcome the mismatching effect of textureless and texture-repeated areas [77], [78], [79]. However, such methods tend to learn prior knowledge of depth estimation from training data and produce error depth prediction for the real-world scenarios that are obviously different from training scenarios.

In this article, we comprehensively apply active light sensing technologies and passive depth estimation methods in V-Classroom. Sparse consumer-grade RGBD cameras are configured in the physical-world classroom for capturing the prior depth information. Then, the image-based depth completion is employed to refine the depth maps collected by the sensors, to overcome the false and empty depth values affected by illumination, black objects, transparent materials, and so on.

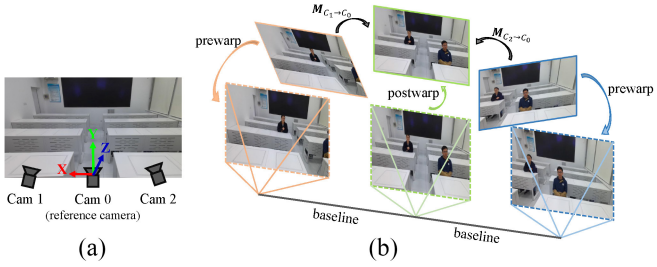


Fig. 3. (a) Implementation example of the camera setup in the physical-world classroom with the V-Classroom's local coordinate system. (b) Color images of the physical-world classroom with learners captured by the three cameras, and coordinate transformations in the V-Classroom system for camera calibration and novel view synthesis.

III. CPSS Paradigm of V-Classroom

The CPSS has been explored in many researches and defined in different perspectives, dependent on the application fields. In our work, we summarize a generic notion and adopt a perspective on CPSS represented by the following definition.

Definition 1: A *Social System* is a system that involves interacting human objects with individual cognition, preferences, motivation, and behaviors.

Definition 2: CPS is an intelligent system that encompasses all systems and subsystems of cyber and physical systems, the components of and the interactions between them, and the integration of computations and physical processes.

Definition 3: In a general sense, *CPSS* is a system that comprises of a CPS defined in *Definition 2* and a social system defined in *Definition 1*.

Based on the definition of CPSS and the characteristics of Edu-Metaverse, a four-layered architecture of CPSS for Edu-Metaverse is illustrated in Fig. 1. The layered CPSS architecture actually can be realized based on the existing protocols of CPSS and Internet of Things (IoT). The cyber layer in the cyber space supports the intelligent data processing and CPSS controlling for smart decision making. Above this layer, there exists a social layer in the social space representing the education-related human society, and below this layer, there exists a physical layer in the physical space containing physical-world educational scenes, activities with various components. The physical layer and the cyber layer are connected by the sensor and other actuator networks, and the humans in social layer participate in the CPSS system operation through VR agents or certain multimedia devices connected to the network. Additionally, the virtual layer denotes the software-defined virtual world developed by executing a virtualization process of both physical world and human society.

To concretize the layered CPSS architecture, we concentrate on the remote teaching activities and accordingly design the V-Classroom, a lecturer-centered consumer-grade immersive teaching system. The overall framework and workflow of V-Classroom deployed as a substructure of CPSS-for-Edu-Metaverse architecture is illustrated in Fig. 2. The V-Classroom system, defined in a CPSS paradigm, encompasses a classroom side in the physical world, a lecturer side in the social world, and a cloud side in the cyber world. Such system, which focuses on the characteristic of immersion, can be regarded as a supporting part for Edu-Metaverse. Particularly, we take the lecturers' factors, such as belonging sense to the classroom, teaching motivation and enthusiasm, and social factors, such as affordability, reproducibility, flexibility for education equity, into consideration, which results in the two special advantages of V-Classroom. First, V-Classroom is a real-time 6DoF video communication system specially for the lecturers, so that the lecturers' space immersion and the mental experience can be improved. Second, V-Classroom workflow is standardized by means of a CPSS paradigm, which enables reducing costs, improving reproducibility, and achieving off-the-shelf and flexible system construction. It is conducive to the generalization of such system in different regions, thus, promoting the education equity. The CPSS deployment and overall framework of V-Classroom consisting of three parts is described as follows.

Physical World: In the physical world, we start by defining the V-Classroom's local coordinate system and calibrating the intrinsic/extrinsic parameters of all the RGBD cameras [80]. An implementation example of the camera setup in the physical-world classroom with the V-Classroom's local coordinate system is shown in Fig. 3(a). The center camera is regarded as the reference camera along with the optical center defined as the origin of the local coordinate system. Then, we define $\mathbf{X}$ direction as the horizontal direction of the front blackboard in the physical-world classroom and $\mathbf{Y}$ direction as the upward direction that is perpendicular to the floor plane. The $\mathbf{Z}$ direction is determined by $\mathbf{X} \times \mathbf{Y}$ accordingly. The $\mathbf{XZ}$ plane is the floor plane, and the scale of this local coordinate system is set to be the same as the scale of the physical world. Based on this setup, the coordinate transformations in the V-Classroom system for camera calibration and novel view synthesis can be clear, as shown in Fig. 3(b). Based on the collected RGBD data of static classroom and the camera intrinsics/extrinsics, a prior geometric model can be obtained through MBR-related technologies, such as point cloud registration and fusion.

Cyber World: In the cyber world of V-Classroom, the collected RGB and processed depth frames are encoded and transmitted to the cloud side by means of standard video compression technologies and communication protocols. The data flow of multiple V-Classroom systems are distributed and managed intelligently in the cloud. Then, the data are decoded and

transmitted to local server deployed with an SDK including our developed V-Classroom algorithms for the intelligent data processing and view synthesis. Our current system focuses on realistic classroom video rendering to establish visual immersion for remote lecturers. For audio transmission, conventional equipment and communication software can be used. In practice, we observe that the video delay is almost the same as the audio transmission delay so no special processing is applied.

Social World: In the social world, our current system focuses on lecturers' mental experience of visual immersion. To achieve an immersive user experience, real-time 3-D displaying technologies, such as wearable VR or augmented reality (AR) devices, are considered as the main interaction interface. Therefore, we provide each V-Classroom lecturer with a physical VR device for sending 6DoF view-change messages and showing view-dependent content of the remote classroom along with learner characters. When the V-Classroom lecturer wears a given VR device, the lecturer's position will be initialized as the origin of the local coordinate system of V-Classroom and the frame at corresponding view will be displayed. Of course, the V-Classroom lecturers can set their preferred initial positions freely. Our system will perform proper coordinate transformations instantaneously. Once the initial position of the lecturer is determined in the local coordinate system of V-Classroom, the lecturer's position change will be sent through the VR device in the form of a 6DoF parameter and the V-VA will be triggered. The frame at the corresponding viewpoint will be synthesized and sent back to the lecturer side to display in a precollected screen.

IV. Key Algorithms of V-Classroom

The core objective of V-Classroom is to achieve real-time rendering of remote classroom scenarios with consumer-grade hardware setup in the physical world. When designing a rendering algorithm for V-Classroom we face three challenges.

1) We have to deal with wide baselines in rendering. Due to the large classroom sizes and the sparse cameras, the view difference between adjacent cameras are large and the rendering must cover a wide range of virtual viewpoints.
2) The V-Classroom is required to display high-definition videos for lecturers wearing a VR device, where rendering flaws can be easily noticed by users. Thus, the synthesized virtual view images should be visually comfortable.
3) The view synthesis algorithm must run at real-time to ensure the whole V-Classroom system functions well.

However, we are not aware of any novel view synthesis method that can perform high-quality rendering in case of wide baselines. Most existing methods cannot achieve online capturing and high-quality rendering in real time, while other real-time solutions suffer from severe artifacts, producing incomplete regions or only synthesizing low-resolution results.

To address the issues, we develop the V-Classroom algorithms, including a V-DA and a V-VA, based on a few key insights. First, we leverage RGBD cameras and acquire prior depth maps for static classrooms to ease the burden of geometry estimation. The acquired depth maps, although quite noisy and cannot be directly used, can provide reasonable depth priors for V-DA and V-VA process. And we execute a depth optimization and completion technology, considering the characteristics of classroom scenes, in V-DA to obtain high-quality and dense depth maps, which leads to improved rendering quality especially for our wide-baseline scenarios. Second, we incorporate the state-of-the-art long-time video object segmentation technologies in V-VA for preliminary processing the learner characters rendering, so as to improve the visual quality and robustness of novel view synthesis of the whole dynamic classroom. Finally, we have applied parallel computing and GPU acceleration as much as possible in the implementation process to ensure the whole system run in real time. The details are described as follows.

A. V-Classroom Depth Algorithm

V-DA is implemented for static classroom scenarios and includes three stages: 1) prior depth acquisition; 2) image-guided depth completion; and 3) planar-constrained depth optimization. The workflow and intermediate results are shown in Fig. 4(b).

Prior Depth Acquisition: Inevitably, there exist false and empty depth values in the initial depth maps collected by RGBD sensors, due to the illumination, black surface, glass materials, and other challenging factors in the physical-world classroom. Thus, we concentrate on the optimization and completion of depth maps. First, we apply a point cloud registration and fusion toolkit to generate a prior geometric model in the form of 3-D point cloud from the camera-collected images through preliminary camera calibration and projection transformation. Then, the prior depth map D_{C_i} of the reference view C_i can be obtained by projecting the prior geometric model to a given position using the camera calibration results. Actually, this step plays a role as initial depth refinement aided by multiview depth information. Such prior depth maps are fed into the subsequent depth completion and optimization stages for improving the quality and robustness of V-DA.

Image-Guided Depth Completion: There still exist many empty depth values in the prior depth maps D_{C_i}. We further explore an image-guided depth completion for obtaining dense and intact depth maps D'_{C_i}. Given a reference-view image $I_{C_i} \in \mathbb{R}^{W\times H}$ with corresponding prior depth map $D_{C_i} \in \mathbb{R}^{W\times H}$, we need to find $\hat{f}$ that approximates a true function $f : \mathbb{R}^{W\times H} \times \mathbb{R}^{W\times H} \to \mathbb{R}^{W\times H}$ where $f(I_{C_i}, D_{C_i}) = D'_{C_i}$. The problem can be formulated as

$$\text{argmin}_{\hat{f}} \left\| \hat{f}\left(I_{C_i}, D_{C_i}\right) - f\left(I_{C_i}, D_{C_i}\right) \right\|. \tag{1}$$

In this stage, we realize $\hat{f}$ via a series of image processing operations [81], which can be achieved by the following steps in turn: depth inversion with $D_{\text{inverted}} = 10\text{-D}$ for setting a buffer (2m) between effective depth and null value, dilation with custom diamond kernel, small hole closure and fill, large hole fill, Median and Gaussian blur for smoothing local planes and edges, and depth inversion for restoring invert depth. Additionally, we incorporate the image colorization techniques based on clustering and distance transformation for null value

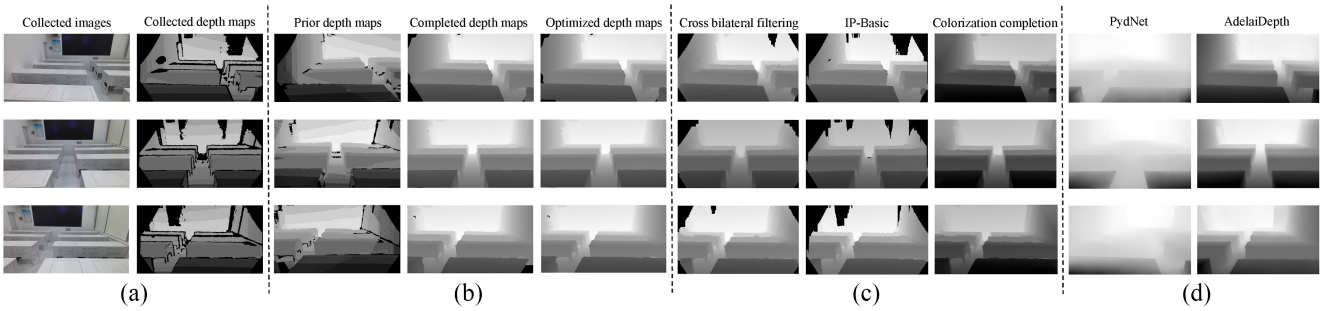


Fig. 4. V-DA workflow and visual comparison results. (a) Color images and corresponding depth maps collected by three RGBD cameras. (b) Depth map results of each stage of our V-DA workflow, including the prior depth map D_{C_i} acquired from the prior geometric model, the image guided depth completion results D'_{C_i}, and the planar constrained depth optimization results D''_{C_i}. (c) Depth maps resulted from other representative depth completion methods. (d) Depth maps resulted from other representative passive depth estimation methods.

completion. That is, if two adjacent pixels r and s have similar intensities, they should have similar colors and similar depth values. Thus, we minimize the distance between the color $I_{C_i}(r)$ at r and the weighted average of the colors at neighboring pixels, which ensures minimizing the distance between the depths of neighboring pixels

$$\min J(I_{C_i}) = \sum_r \left(I_{C_i}(r) - \sum_{s \in N(r)} w_{rs} I_{C_i}(s) \right)^2. \tag{2}$$

Planar-Constrained Depth Optimization: Particularly, there are a lot of structural features (line segments, planes, etc.) and geometric constraints (planarity, orthogonality, etc.) in the classroom scenes. Therefore, we explore a planar constrained depth optimization method to improve the depth accuracy and integrity of the coplanar regions. In the local coordinate system of V-Classroom, a plane equation can be expressed as $aX + bY + cZ + d = 0$. Mark the existing 3-D points $q = [u, v, Z]$ of certain plane in D'_{C_i}, where (u, v) denotes the pixel coordinate of q, and Z denotes the depth value of q querying from D'_{C_i}. The pixel coordinates can be transformed into the local coordinate of V-Classroom through calibrated camera intrinsics, that is, $q' = [X, Y, Z]$, where

$$X = \frac{Z(u - u_0)}{f_x}, \quad Y = \frac{Z(v - v_0)}{f_y}. \tag{3}$$

u_0 and v_0 denotes the origin of pixel coordinate system and f_x and f_y denotes the focal length of camera. By means of all the marked q, the parameters of corresponding plane equation can be solved using the least square method. Obviously, the planarity of objects remains in multiview images, so does in depth maps. Thus, we exploit this property to optimize the D'_{C_i}. Taking the center camera as reference to mark the reference plane equation $a_0X + b_0Y + c_0Z + d_0 = 0$ in D'_{C_0} and denoting the calibrated extrinsics of other cameras as $\mathbf{P}_{Ci} = [\mathbf{R}_{C_i} | \mathbf{T}_{C_i}]$, the transformed plane equation in other-view depth map D'_{C_i} is calculated by $a_iX + b_iY + c_iZ + d_i = 0$, $[a_i, b_i, c_i]^T = [a_0, b_0, c_0]^T \times \mathbf{R}_{C_i}$ and $d_i = d_0 + [a_0, b_0, c_0]^T \times \mathbf{T}_{C_i}$. Finally, the optimized depth map D''_{C_i} can be obtained with the optimized depth value Z'_{C_i} meeting the following equation:

$$\left[X'_{C_i}, Y'_{C_i}, Z'_{C_i}\right]^T = \left[X_{C_i}, Y_{C_i}, Z_{C_i}\right]^T \times \mathbf{R}_{C_i} + \mathbf{T}_{C_i}. \tag{4}$$

B. V-Classroom View Algorithm

In the SDK of our system, the input video frames with the resulted depth maps from V-DA are fed into the V-VA, which is implemented for the dynamic classroom scenarios with learners. V-VA is designed to be composed of learner characters segmentation and virtual view synthesis.

Learner Characters Segmentation: In order to avoid the failed warping and distorted rendering of learner character caused by the depth error and flexible human body, we apply the video object segmentation technologies as a preliminary stage for processing the learner characters rendering, so as to improve the visual quality and robustness of novel virtual view synthesis. Considering the time requirements of the teaching process, we explore a state-of-the-art long-time video object segmentation architecture named XMem [82] that incorporates multiple independent yet deeply connected feature memory stores, including a rapidly updated sensory memory, a high-resolution working memory, and a compact, thus, sustained long-term memory. The first frame is used to initialize different characteristic memory pools, and the XMem tracks multiple character targets and generates corresponding mask maps for each subsequent frame. Then, we update each characteristic storage pool with different frequencies. The sensory memory is updated every frame, and the working memory is updated once per interval r frame. Also, the working memory is consolidated into the long-term memory in a compact form when it is full, and the long-term memory will forget obsolete features over time. The mask of learner characters is computed via the XMem decoder, where the input is the short-term sensory memory $\mathbf{h}_{t-1} \in \mathbb{R}^{C^h \times H \times W}$ and the feature $\mathbf{F} \in \mathbb{R}^{C^v \times H \times W}$ representing information stored in both long-term and working memory. Following [82], we perform the model training on public video object segmentation datasets of YouTubeVOS and DAVIS. Then, the trained model is directly applied in the V-Classroom for the learner characters segmentation, without finetune in our classroom scenes.

Virtual View Synthesis: In this stage, we process the view synthesis of static classroom scenes and the learners characters rendering separately. Then they are blended into the final synthesized results at a given virtual viewpoint V_i. First, the virtual-view depth map D_{V_i} of static classroom scenes is obtained through 3-D projection transformation based on

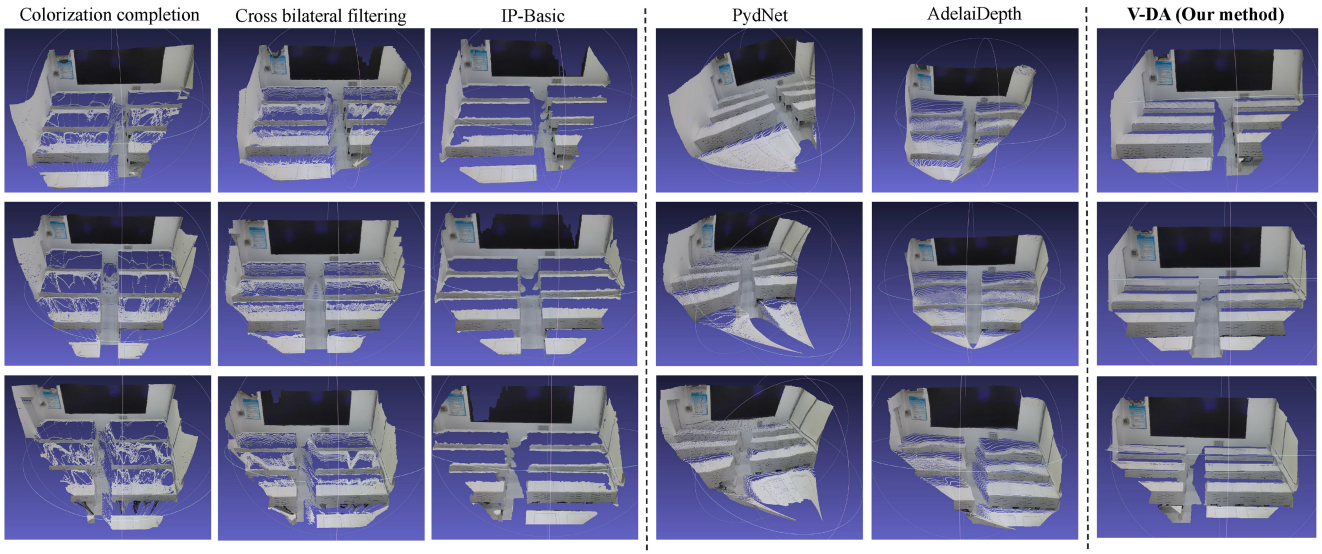


Fig. 5. Qualitative comparison in the form of 3-D point cloud between our V-DA and other representative depth completion/estimation methods.

the collected frames I_{C_i} and resulted depth maps D'_{C_i} from V-DA. Second, for the novel view synthesis of static classroom scenes, we warp the I_{C_i} to the virtual view as

$$I^s_{C_i \to V_i} = \mathbf{warp}\big(I_{C_i} | D_{V_i}\big). \tag{5}$$

The warping function is computed based on the 6DoF view-change parameters sent from the VR device of the lecturer and the calibrated camera intrinsics and extrinsics. And a multiview texture fusion is obtained by weighted averaging the warped frames as

$$I^s_{V_i} = \frac{\sum_i M_{V_i} \cdot I^s_{C_i \to V_i}}{\sum_i M_{V_i}} \tag{6}$$

where M_{V_i} denotes the visibility mask of $I^s_{C_i \to V_i}$ according to the depth comparison and the calculation is done pixelwise. Third, the learner characters are warped to the virtual view based on the segmentation mask and initial depth maps, similar as the warping process of static classroom scenes. Finally, the $I^s_{V_i}$ and the learner characters rendering result are fused to produce the final novel-view result I_{V_i}. Note that in this process, occlusions in the classroom scene should be determined through a depth-based Z-buffer algorithm, so as to perform correct fusion. Additionally, a filtering-based hole filling technology is adopted for optimizing final view synthesis results.

V. Experiments and Results

A. Data Acquisition and Implementation Details

Data Acquisition: We implement two V-Classroom instances to evaluate our system in different classroom scenarios, including a multimedia classroom and a seminar classroom in the Artificial Intelligence College of Beijing Normal University. The two classrooms are located in one building and connected in a LAN network with 1Gb/s bandwidth. We set up three consumer-grade RGBD sensors, Kinect Azure, in the physical-world classroom scenes. The multimedia classroom has a size of 6 m × 5 m × 2.4 m with four rows of desks and the three sensors are placed horizontally with a 80 cm-baseline between adjacent cameras The seminar classroom has a size of 4.5 m × 5 m × 2.8 m with two rows of desks and the three sensors are placed horizontally with a 40 cm-baseline between adjacent cameras.

Implementation Details: In the physical world of V-Classroom, based on the collected RGBD data of static classroom and the camera intrinsics K_{C_i} provided by AzureKinect SDK, we register and fuse projected 3-D point clouds using the Meshlab software, where the camera extrinsics P_{C_i} in the local coordinate system of V-Classroom can be obtained. In the cyber world of V-Classroom, the RGB and depth frames are encoded into JPEG images separately and transmited via the TCP/IP protocol to the cloud side. Then, the data are decoded and transmitted to local server, where the SDK is deployed on a consumer-grade PC with Intel Core i7-10700 CPU, 16GB memory, and one NVIDIA GeForce RTX 2060 Super GPU. The V-Classroom lecturer can choose to wear a NOLO X1 4K VR device or use a mouse-interacted interface. Based on the 6DoF view-change messages sent by the lecturer side, the frame at the corresponding viewpoint, with the resolution of 960×540, is synthesized by triggering the V-VA and sent back to the lecturer side to display in a precollected screen. Our current system can achieve 15 frames/s on one NVIDIA GeForce RTX 2060 Super GPU and the network delay is admissible for the teaching scenario without noticeable discomfort.

B. V-Classroom Depth Evaluation

To evaluate the V-DA performance of our method, we conduct qualitative evaluations on our collected classroom scenarios, as shown in Figs. 4 and 5.

Fig. 4 demonstrates the depth map results of each stage of V-DA workflow as well as some visual comparisons with other representative depth completion methods and depth estimation methods. Fig. 4(a) shows the color images and

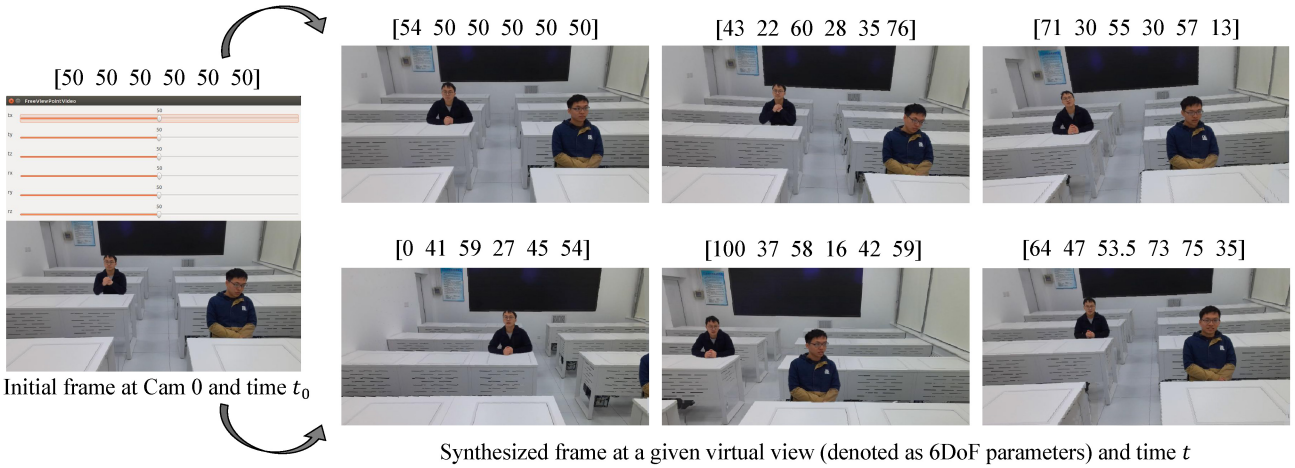


Fig. 6. Some visual results of virtual view synthesis of V-Classroom with a mouse-interacted interface.

corresponding depth maps collected by three RGBD cameras. And the processed depth maps with intermediate results of our V-DA are shown in Fig. 4(b), including the prior depth map D_{C_i} acquired from the prior geometric model, the image-guided depth completion results D'_{C_i}, and the final depth optimization results D''_{C_i} through planar constraint. In Fig. 4(c), we provide the resulted depth maps from some representative depth completion methods [81], [83], where the inputs are the collected depth maps. And in Fig. 4(d), we provide the resulted depth maps from some representative passive depth estimation methods [84], [85], where the inputs are the collected images. It can be observed that our V-DA achieves more complete depth maps along with more reasonable distribution.

Although there are no ground-truth depth maps for a quantitative evaluation of depth accuracy, we perform a qualitative comparison to evaluate the accuracy and rationality of the depth maps resulted from our V-DA intuitively. That is, we generate a 3-D point cloud for each depth map result along with the collected image through 3-D projection, as shown in Fig. 5. It can be observed that our V-DA achieves more accurate depth results with few noise points and reasonable plane geometry.

C. V-Classroom View Evaluation

To evaluate the V-VA performance of our method, we conduct qualitative evaluations on our collected classroom scenarios, as shown in Figs. 6–8.

As shown in Fig. 6, we additionally design a mouse-interacted interface to demonstrate the virtual view synthesis results of V-Classroom. We first define the initial position of a V-Classroom lecturer as the origin of local coordinate system of V-Classroom (i.e., Cam 0). We denote this initial position as 6DoF parameters [50, 50, 50, 50, 50, 50] and the initial time as t_0. Then, the synthesized frame at a given virtual view (denoted as 6DoF parameters above the frame) and arbitrary time t. If a V-Classroom lecturer is not used for wearing any VR devices, such mouse-interacted interface will be a good choice.

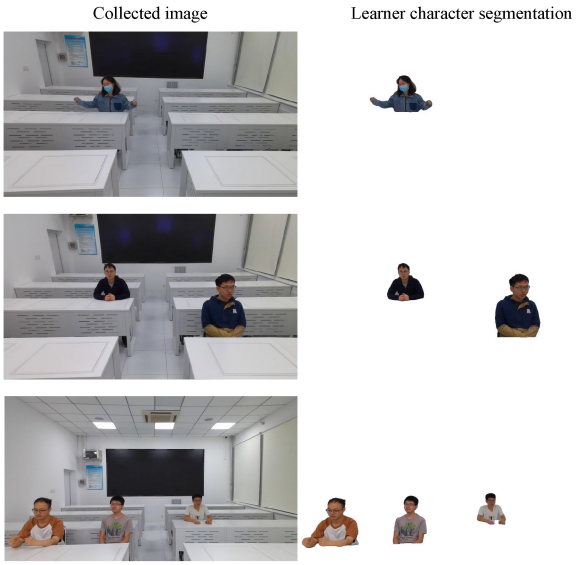


Fig. 7. Learner characters segmentation results of classroom scenarios containing single learner (first row), two learners (second row), and three learners (third row).

Fig. 7 demonstrates the learner characters segmentation results of classroom scenarios containing single learner, two learners, and three learners. It can be observed that our method can achieve visually reasonable and comfortable segmentation results, even for the learners whose clothing color is very close to the background or whose scale is small.

Additionally, we present some visual comparison between our method and other depth-based view synthesis methods, as shown in Fig. 8. Both Figs. 6 and 8 demonstrate the effectiveness of V-VA. Our V-Classroom algorithms can produce visually reasonable and comfortable virtual-view results and achieve significantly higher visual quality compared with DIBR-related methods.

D. Ablation Studies

We conduct the ablation studies of different architecture designing choices of V-Classroom algorithms. Though there

Fig. 8. Visual comparison between our method and other depth-based view synthesis methods.

TABLE I
ABLATION STUDIES OF THE V-DA AND V-VA ON OUR COLLECTED VIDEO DATA IN THE MULTIMEDIA CLASSROOM. AVG., MAX., AND MIN. DENOTE THE AVERAGE VALUE, MAXIMUM VALUE, AND MINIMUM VALUE OF THE EVALUATION RESULTS OF COLLECTED 300 FRAMES, RESPECTIVELY. W/O DENOTES WITHOUT. SEG. AND FUS. DENOTE THE LEARNER CHARACTER SEGMENTATION STAGE AND MULTIVIEW TEXTURE FUSION STAGE, RESPECTIVELY

Methods	Image PSNR (↑)			Image SSIM (↑)		
	Avg.	Max.	Min.	Avg.	Max.	Min.
Ours w/o V-DA	18.54	18.97	17.76	69.75	70.52	69.14
Ours w/o seg. of V-VA	20.98	21.02	20.55	86.46	86.71	86.30
Ours w/o fus. of V-VA	22.76	22.78	22.64	87.92	87.95	87.91
Ours	**25.88**	**25.98**	**25.75**	**88.28**	**88.43**	**88.10**

are no ground-truth virtual-view results for a quantitative evaluation of V-Classroom, we specially consider a quantitative evaluation method that exploits our data of 300 frames of the multimedia classroom and 300 frames of the seminar classroom, collected with three cameras. We adopt different methods to synthesize the virtual-view frame at Cam 0 utilizing the frames collected by the remaining two cameras as reference views. Then, we calculate the peak signal to noise ratio (PSNR) and structural similarity index (SSIM) between the synthesized frame and collected frame at Cam 0, at the same point in time. The results of two different classroom scenarios, including the multimedia classroom and the seminar classroom, are shown in Tables I and II. We can observe that the V-DA and V-VA consistently increase the accuracy of our architecture.

VI. CONCLUSION

In this article, we explore the application of cyber–physical–social intelligence in Edu-Metaverse and construct a layered CPSS architecture for Edu-Metaverse, taking the social and lecturers' factors into consideration. Based on it, we focus on the lecturers' mental experience in remote teaching activity and specially design a lecturer-centered consumer-grade immersive teaching system, named V-Classroom. A CPSS paradigm of V-Classroom is first introduced to standardize and simplify the workflow. Furthermore, to achieve real-time rendering of classroom scenarios with consumer-grade hardware setup, we propose the V-Classroom algorithms including V-DA and V-VA. And the experiments on two different classroom scenarios demonstrate the effectiveness of our proposed method. We believe there is much more to be discovered along this direction and V-Classroom will inspire more intelligent teaching technologies on top of our work.

TABLE II
ABLATION STUDIES OF THE V-DA AND V-VA ON OUR COLLECTED VIDEO DATA IN THE SEMINAR CLASSROOM. AVG., MAX., AND MIN. DENOTE THE AVERAGE VALUE, MAXIMUM VALUE, AND MINIMUM VALUE OF THE EVALUATION RESULTS OF COLLECTED 300 FRAMES, RESPECTIVELY. W/O DENOTES WITHOUT. SEG. AND FUS. DENOTE THE LEARNER CHARACTER SEGMENTATION STAGE AND MULTIVIEW TEXTURE FUSION STAGE, RESPECTIVELY

Methods	Image PSNR (↑)			Image SSIM (↑)		
	Avg.	Max.	Min.	Avg.	Max.	Min.
Ours w/o V-DA	19.34	19.89	18.84	74.63	75.39	74.11
Ours w/o seg. of V-VA	22.68	23.10	22.26	84.72	85.15	84.47
Ours w/o fus. of V-VA	24.16	24.28	24.07	84.94	85.19	84.84
Ours	**25.48**	**26.18**	**24.79**	**86.32**	**86.74**	**86.07**

REFERENCES

[1] V. T. Le N. H. Nguyen, T. L. N. Tran, L. T. Nguyen, T. A. Nguyen, and M. T. Nguyen, "The interaction patterns of pandemic-initiated online teaching: How teachers adapted," *System*, vol. 105, Apr. 2022, Art. no. 102755.

[2] I. Chirikov, T. Semenova, N. Maloshonok, E. Bettinger, and R. F. Kizilcec, "Online education platforms scale college STEM instruction with equivalent learning outcomes at lower cost," *Sci. Adv.*, vol. 6, no. 15, p. eaay5324, 2020.

[3] B. B. Lockee, "Online education in the post-COVID era," *Nat. Electron.*, vol. 4, no. 1, pp. 5–6, 2021.

[4] F.-Y. Wang, "The DAO to MetaControl for MetaSystems in metaverses: The system of parallel control systems for knowledge automation and control intelligence in CPSS," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 11, pp. 1899–1908, Nov. 2022.

[5] F.-Y. Wang, "Metavehicles in the metaverse: Moving to a new phase for intelligent vehicles and smart mobility," *IEEE Trans. Intell. Veh.*, vol. 7, no. 1, pp. 1–5, Mar. 2022.

[6] A. Song, W.-N. Chen, T. Gu, H. Yuan, S. Kwong, and J. Zhang, "Distributed virtual network embedding system with historical archives and set-based particle swarm optimization," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 2, pp. 927–942, Feb. 2021.

[7] J. Leng et al., "Blockchain-secured smart manufacturing in industry 4.0: A survey," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 237–252, Jan. 2021.

[8] L. Zou, Z. Wang, Q.-L. Han, and D. Zhou, "Moving horizon estimation of networked nonlinear systems with random access protocol," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 5, pp. 2937–2948, May 2021.

[9] X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F.-Y. Wang, "From features engineering to scenarios engineering for trustworthy AI: I&i, C&C, and V&V," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 18–26, Jul./Aug. 2022.

[10] X. Wang, J. Yang, J. Han, W. Wang, and F.-Y. Wang, "Metaverses and DeMetaverses: From digital twins in CPS to parallel intelligence in CPSS," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 97–102, Jul./Aug. 2022.

[11] M. Wang, H. Yu, Z. Bell, and X. Chu, "Constructing an edu-metaverse ecosystem: A new and innovative framework," *IEEE Trans. Learn. Technol.*, early access, Sep. 29, 2022, doi: 10.1109/TLT.2022.3210828.

[12] J. Wu and G. Gao, "Edu-metaverse: Internet education form with fusion of virtual and reality," in *Proc. Int. Conf. Humanities Soc. Sci. Res.*, 2022, pp. 1082–1085.

[13] L. Buchan, M. Hejmadi, L. Abrahams, and L. D. Hurst, "A RCT for assessment of active human-centred learning finds teacher-centric non-human teaching of evolution optimal," *NPJ Sci. Learn.*, vol. 5, no. 1, pp. 1–20, 2020.

[14] F.-Y. Wang, "The emergence of intelligent enterprises: From CPS to CPSS," *IEEE Intell. Syst.*, vol. 25, no. 4, pp. 85–88, Jul./Aug. 2010.

[15] Y. Zhao, Z. Chen, C. Zhou, Y.-C. Tian, and Y. Qin, "Passivity-based robust control against quantified false data injection attacks in cyber-physical systems," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 8, pp. 1440–1450, Aug. 2021.

[16] Y. Wu and J. Dong, "Cyber-physical attacks against state estimators based on a finite frequency approach," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 2, pp. 864–874, Feb. 2021.

[17] D. Ding, Q.-L. Han, X. Ge, and J. Wang, "Secure state estimation and control of cyber-physical systems: A survey," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 176–190, Jan. 2021.

[18] Z. Zhou, B. Wang, M. Dong, and K. Ota, "Secure and efficient vehicle-to-grid energy trading in cyber physical systems: Integration of blockchain and edge computing," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 1, pp. 43–57, Jan. 2020.

[19] J. Lai, X. Lu, X. Yu, A. Monti, and H. Zhou, "Distributed voltage regulation for cyber-physical microgrids with coupling delays and slow switching topologies," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 1, pp. 100–110, Jan. 2020.

[20] Y. B. Aissa, A. Bachir, M. Khalgui, A. Koubaa, Z. Li, and T. Qu, "On feasibility of multichannel reconfigurable wireless sensor networks under real-time and energy constraints," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 3, pp. 1446–1461, Mar. 2021.

[21] A. O. Akmandor, X. Dai, and N. K. Jha, "YSUY: Your Smartphone understands you—Using machine learning to address fundamental human needs," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 12, pp. 7553–7568, Dec. 2021.

[22] Y. Ren and G.-P. Li, "An interactive and adaptive learning cyber physical human system for manufacturing with a case study in worker machine interactions," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 6723–6732, Oct. 2022.

[23] P. Bhandari, C. Boyle, J. Gong, K. M. Y. Law, and D. Creighton, "Ongoing transformation of critical infrastructure systems as cyberphysical-human systems," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2021, pp. 3342–3347.

[24] A. Macías and E. Navarro, "Paradigms for the conceptualization of cyber-physical-social-thinking hyperspace: A thematic synthesis," *J. Ambient Intell. Smart Environ.*, vol. 14, no. 4, pp. 285–316, 2022.

[25] H. Ning et al., "Cyberology: Cyber-physical-social-thinking spaces based discipline and inter-discipline hierarchy for metaverse (general cyberspace)," *IEEE Internet Things J.*, early access, Oct. 28, 2022, doi: 10.1109/JIOT.2022.3217821.

[26] K. Rijswijk et al., "Digital transformation of agriculture and rural areas: A socio-cyber-physical system framework to support responsibilisation," *J. Rural Stud.*, vol. 85, pp. 79–90, Jul. 2021.

[27] M. A. Hamzaoui and N. Julien, "Social cyber-physical systems and digital twins networks: A perspective about the future digital twin ecosystems," *IFAC-PapersOnLine*, vol. 55, no. 8, pp. 31–36, 2022.

[28] S. A. Barkalov, M. I. Lomakin, L. E. Mistrov, V. P. Morozov, and O. I. Zakharova, "Information support of decision making in social-cyber-physical systems of machine-building production based on ontological model of knowledge representation," in *Proc. AIP Conf.*, 2021, Art. no. 40018.

[29] B. B. Gupta, K.-C. Li, V. C. Leung, K. E. Psannis, and S. Yamaguchi, "Blockchain-assisted secure fine-grained searchable encryption for a cloud-based healthcare cyber-physical system," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 12, pp. 1877–1890, Dec. 2021.

[30] A. White, A. Karimoddini, and M. Karimadini, "Resilient fault diagnosis under imperfect observations—A need for industry 4.0 era," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 5, pp. 1279–1288, Sep. 2020.

[31] J. Leng et al., "ManuChain: Combining permissioned blockchain with a holistic optimization model as bi-level intelligence for smart manufacturing," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 1, pp. 182–192, Jan. 2020.

[32] T. Liu, B. Tian, Y. Ai, and F.-Y. Wang, "Parallel reinforcement learning-based energy efficiency improvement for a cyber-physical system," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 2, pp. 617–626, Mar. 2020.

[33] X. Zhao, S. Zou, and Z. Ma, "Decentralized resilient H_∞ load frequency control for cyber-physical power systems under DoS attacks," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 11, pp. 1737–1751, Nov. 2021.

[34] H. Sun, C. Peng, D. Yue, Y. L. Wang, and T. Zhang, "Resilient load frequency control of cyber-physical power systems under QoS-dependent event-triggered communication," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 4, pp. 2113–2122, Apr. 2021.

[35] S. Talukder, M. Ibrahim, and R. Kumar, "Resilience indices for power/cyberphysical systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 4, pp. 2159–2172, Apr. 2021.

[36] M. Al-Sharman et al., "A sensorless state estimation for a safety-oriented cyber-physical system in urban driving: Deep learning approach," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 1, pp. 169–178, Jan. 2021.

[37] S. Han et al., "From software-defined vehicles to self-driving vehicles: A report on CPSS-based parallel driving," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 1, pp. 6–14, Oct. 2018.

[38] F.-Y. Wang, N.-N. Zheng, D. Cao, C. M. Martinez, L. Li, and T. Liu, "Parallel driving in CPSS: A unified approach for transport automation and vehicle intelligence," *IEEE/CAA J. Automatica Sinica*, vol. 4, no. 4, pp. 577–587, Sep. 2017.

[39] A. E. Leonova, V. I. Karpov, Y. Y. Chernyy, and E. V. Romanova, "Transformation PLM-systems into the cyber-physical systems for the information provision for Enterprise management," in *Proc. Int. Conf. Cyber-Phys. Syst. Control*, 2019, pp. 431–439.

[40] Z. Liu, D.-S. Yang, D. Wen, W.-M. Zhang, and W. Mao, "Cyber-physical-social systems for command and control," *IEEE Intell. Syst.*, vol. 26, no. 4, pp. 92–96, Jul./Aug. 2011.

[41] C. Zhao, Y. Lv, J. Jin, Y. Tian, J. Wang, and F.-Y. Wang, "DeCAST in TransVerse for parallel intelligent transportation systems and smart cities: Three decades and beyond," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 6, pp. 6–17, Nov./Dec. 2022.

[42] G. Xiong et al., "Cyber-physical-social systems for smart city: An implementation based on intelligent loop," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 501–506, 2020.

[43] T. Roy, A. Tariq, and S. Dey, "A socio-technical approach for resilient connected transportation systems in smart cities," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5019–5028, Jun. 2022.

[44] C. Dede, "Immersive interfaces for engagement and learning," *Science*, vol. 323, no. 5910, pp. 66–69, 2009.

[45] S. Cai, X. Jiao, and B. Song, "Open another door to education—Applications, challenges and perspectives of the educational metaverse," *Metaverse*, vol. 3, no. 1, p. 12, 2022.

[46] K. Getchell, I. Oliver, A. Miller, and C. Allison, "Metaverses as a platform for game based learning," in *Proc. IEEE Int. Conf. Adv. Inf. Netw. Appl.*, 2010, pp. 1195–1202.

[47] S. H. Damas and M. J. B. de Gracia, "Immersive journalism: Advantages, disadvantages and challenges from the perspective of experts," *J. Media*, vol. 3, no. 2, pp. 330–347, 2022.

[48] A. Jovanović and A. Milosavljević, "VoRtex metaverse platform for gamified collaborative learning," *Electronics*, vol. 11, no. 3, p. 317, 2022.

[49] H. Lee and Y. Hwang, "Technology-enhanced education through VR-making and metaverse-linking to foster teacher readiness and sustainable learning," *Sustainability*, vol. 14, no. 8, p. 4786, 2022.

[50] S. Park and S. Kim, "Identifying world types to deliver gameful experiences for sustainable learning in the metaverse," *Sustainability*, vol. 14, no. 3, p. 1361, 2022.

[51] V. M. Petrović and B. D. Kovačević, "AViLab—Gamified virtual educational tool for introduction to agent theory fundamentals," *Electronics*, vol. 11, no. 3, p. 344, 2022.

[52] H. Duan, J. Li, S. Fan, Z. Lin, X. Wu, and W. Cai, "Metaverse for social good: A university campus prototype," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 153–161.

[53] "U.K.'s first virtual meeting space from Saïd Business School, University of Oxford." Saïd Business School. Jul. 2018. [Online]. Available: https://www.b4-business.com/article/uks-first-virtual-meeting-space-said-business-school-university-oxford/

[54] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," in *Proc. ACM Siggr. Papers*, 2006, pp. 835–846.

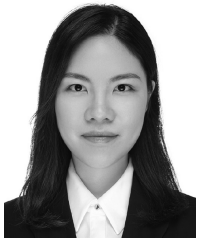
[55] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[56] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, "Dense 3D object reconstruction from a single depth view," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2820–2834, Dec. 2019.

[57] G. Riegler, A. O. Ulusoy, and A. Geiger, "Octnet: Learning deep 3D representations at high resolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3577–3586.

[58] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 605–613.

[59] C. Nash, Y. Ganin, S. A. Eslami, and P. Battaglia, "PolyGen: An autoregressive generative model of 3D meshes," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7220–7229.
[60] A. Luo, T. Li, W.-H. Zhang, and T. S. Lee, "SurfGen: Adversarial 3D shape synthesis with explicit surface discriminators," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16238–16248.
[61] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19313–19322.
[62] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow, "Deep blending for free-viewpoint image-based rendering," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–15, 2018.
[63] J. Pearson, M. Brookes, and P. L. Dragotti, "Plenoptic layer-based modeling for image based rendering," *IEEE Trans. Image Process.*, vol. 22, pp. 3405–3419, 2013.
[64] B. Mildenhall et al., "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–14, 2019.
[65] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.
[66] S. Li, K. Wang, Y. Gao, X. Cai, and M. Ye, "Geometric warping error aware CNN for DIBR oriented view synthesis," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1512–1521.
[67] Y. Ren, B. Liu, R. Cheng, and C. Agia, "Lightweight semantic-aided localization with spinning LiDAR sensor," *IEEE Trans. Intell. Veh.*, early access, Jul. 26, 2021, doi: 10.1109/TIV.2021.3099022.
[68] S. Grollius, M. Ligges, J. Ruskowski, and A. Grabmaier, "Concept of an automotive LiDAR target simulator for direct time-of-flight LiDAR," *IEEE Trans. Intell. Veh.*, early access, Nov. 17, 2021, doi: 10.1109/TIV.2021.3128808.
[69] Y. Li et al., "DELTAR: Depth estimation from a light-weight ToF sensor and RGB image," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 619–636.
[70] M. M. Johari, C. Carta, and F. Fleuret, "DepthInSpace: Exploitation and fusion of multiple video frames for structured-light depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6039–6048.
[71] Y. Zhang and D. L. Lau, "BimodalPS: Causes and corrections for bimodal multi-path in phase-shifting structured light scanners," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 13, 2022, doi: 10.1109/TPAMI.2022.3206265.
[72] D.-Y. Nam and J.-K. Han, "Improved depth estimation algorithm via superpixel segmentation and graph-cut," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2021, pp. 1–7.
[73] G. Van Meerbergen, M. Vergauwen, M. Pollefeys, and L. Van Gool, "A hierarchical symmetric stereo algorithm using dynamic programming," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 275–285, 2002.
[74] Y. Li, R. Yunus, N. Brasch, N. Navab, and F. Tombari, "RGB-D SLAM with structural regularities," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 11581–11587.
[75] H. Xu, Z. Zhou, Y. Qiao, W. Kang, and Q. Wu, "Self-supervised multi-view stereo via effective co-segmentation and data-augmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 3030–3038.
[76] H. Liu, S. Huang, N. Gao, and Z. Zhang, "Binocular stereo vision system based on phase matching," in *Proc. Opt. Metrol. Inspect. Ind. Appl. IV*, 2016, pp. 130–138.
[77] C. Zhou, Y. Liu, Q. Sun, and P. Lasang, "Vehicle detection and disparity estimation using blended stereo images," *IEEE Trans. Intell. Veh.*, vol. 6, no. 4, pp. 690–698, Dec. 2021.
[78] P. Ji, R. Li, B. Bhanu, and Y. Xu, "Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12787–12796.
[79] Y. Zhang, M. Gong, J. Li, M. Zhang, F. Jiang, and H. Zhao, "Self-supervised monocular depth estimation with multiscale perception," *IEEE Trans. Image Process.*, vol. 31, pp. 3251–3266, 2022.
[80] H. Zhang, L. Jin, and C. Ye, "An RGB-D camera based visual positioning system for Assistive navigation by a robotic navigation aid," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 8, pp. 1389–1400, Aug. 2021.
[81] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the CPU," in *Proc. IEEE Conf. Comput. Robot Vis.*, 2018, pp. 16–22.
[82] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 640–658.
[83] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.
[84] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "Towards real-time unsupervised monocular depth estimation on CPU," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 5848–5854.
[85] W. Yin et al., "Towards accurate reconstruction of 3D scene shape from a single monocular image," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 5, 2022, doi: 10.1109/TPAMI.2022.3209968.



Tianyu Shen received the Bachelor of Engineering degree in electronic science and technology and the Bachelor of Management degree in accounting from Xi'an Jiaotong University, Xi'an, China, in 2016, and the Ph.D. degree in social computing from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021.

She is currently a Postdoctoral Research Fellow with the School of Artificial Intelligence, Beijing Normal University, Beijing. Her current research interests include computer vision and pattern recognition.

Shi-Sheng Huang received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2015.

He is currently a Lecturer with the School of Artificial Intelligence, Beijing Normal University, Beijing. He was a Postdoctoral Researcher with Tsinghua University. His primary research interests include fields of computer graphics, computer vision, and visual SLAM.

Deqi Li received the Bachelor of Engineering degree in electronic information engineering and the master's degree in mathematics from China University of Geosciences (Beijing), Beijing, China, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree in computer application technology with the School of Artificial Intelligence, Beijing Normal University, Beijing.

His current research interests include computer vision and pattern recognition.

Zhiyuan Lu received the Bachelor of Science degree in mathematics and applied mathematics from Beijing Normal University, Beijing, China, in 2022, where he is currently pursuing the master's degree in computer application technology with the School of Artificial Intelligence.

His current research interest is computer vision.

Fei-Yue Wang (Fellow, IEEE) received the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined The University of Arizona, Tucson, AZ, USA, in 1990, where he became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center with the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Chinese Talents Program from the State Planning Council, and in 2002, he was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS, and the Vice President of the Institute of Automation, CAS, in 2006. In 2011, he became the State Specially Appointed Expert and the Founding Director of the State Key Laboratory for Management and Control of Complex Systems. He has been the Chief Judge of Intelligent Vehicles Future Challenge since 2009 and the Director of China Intelligent Vehicles Proving Center, Changshu, China, since 2015. He is currently the Director of Intel's International Collaborative Research Institute on Parallel Driving with CAS and Tsinghua University, Beijing. His current research focuses on methods and applications for parallel intelligence, social computing, and knowledge automation.

Dr. Wang received the IEEE ITS Outstanding Application and Research Awards in 2009, 2011, and 2015, respectively, the IEEE SMC Norbert Wiener Award in 2014, and became the IFAC Pavel J. Nowacki Distinguished Lecturer in 2021. In 2007, he received the National Prize in Natural Sciences of China, numerous best papers awards from IEEE TRANSACTIONS, and became an Outstanding Scientist of ACM for his work in intelligent control and social computing. Since 1997, he has been serving as the General or Program Chair of over 30 IEEE, INFORMS, IFAC, ACM, and ASME conferences. He was the President of the IEEE ITS Society from 2005 to 2007, the IEEE Council of RFID from 2019 to 2021, the Chinese Association for Science and Technology, USA, in 2005, the American Zhu Kezhen Education Foundation from 2007 to 2008, the Vice President of the ACM China Council from 2010 to 2011, and the Vice President and the Secretary General of the Chinese Association of Automation from 2008 to 2018. He was the Founding Editor-in-Chief (EiC) of the *International Journal of Intelligent Control and Systems* from 1995 to 2000, *IEEE Intelligent Transportation Systems Magazine* from 2006 to 2007, IEEE/CAA JOURNAL OF AUTOMATICA SINICA from 2014 to 2017, *Journal of Command and Control* (China) from 2015 to 2021, and *Journal of Intelligent Science and Technology* (China) from 2019 to 2021. He was the EiC of the IEEE INTELLIGENT SYSTEMS from 2009 to 2012, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS from 2009 to 2016, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS from 2017 to 2020. He is currently the President of CAA's Supervision Council, the Vice President of IEEE Systems, Man, and Cybernetics Society, and the new EiC of IEEE TRANSACTIONS ON INTELLIGENT VEHICLES. He is a Fellow of INCOSE, IFAC, ASME, and AAAS.

Hua Huang (Senior Member, IEEE) received the B.S. degree in radio technology and the M.S. and Ph.D. degrees in information and communication engineering from Xi'an Jiaotong University, Xi'an, China, in 1996, 2001, and 2006, respectively.

He is currently a Professor with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. His current research interests include image and video processing, computer graphics, and pattern recognition.