Letter

Parallel Vision ⊇ Image Synthesis/Augmentation

Wenwen Zhang, Wenbo Zheng, Qiang Li, and Fei-Yue Wang

Dear Editor,

Scene understanding is an essential task in computer vision. The ultimate objective of scene understanding is to instruct computers to understand and reason about the scenes as humans do. Parallel vision is a research framework that unifies the explanation and perception of dynamic and complex scenes. Parallel vision's rationality has been proven through recent research hotspots in artificial intelligence, like the metaverse, world models, and other concepts. At the same time, the development of modern technology has also provided new technologies and ideas for implementing parallel vision. This letter explores the current status of parallel vision, expands the related research connotation of a parallel vision by using the existing research hotspots, and points out the possible development direction of scene understanding based on parallel vision in the future.

Introduction: Given the potential applications of artificial intelligence (AI) and the rapid development of computing power, researchers from different countries are making great efforts to develop and apply AI to different fields in different aspects. The development of AI in vision, namely computer vision, has achieved unprecedented prosperity, as vision is a crucial method for human beings to obtain information. The ultimate goal of computer vision is to enable agents to first perceive, then understand, and finally interact with the real world like humans, which is scene understanding [1]. To achieve the goal, different vision tasks have been proposed, such as scene recognition, object detection, semantic segmentation, and panoramic segmentation in 2D/3D space. The rapid advancement of deep learning has greatly improved the progress of related tasks in computer vision. With large scale datasets, deep learning based models work well in different vision tasks, some of which even surpass humans in certain aspects [2]. However, in order to train the model effectively, it requires the training set to be an independent identically distribution (i.i.d.) [3]. While it is time-consuming and laborious to collect and label large amounts of data under i.i.d. Meanwhile, it is not possible to verify whether the trained model is effective in real scenes.

Considering these, Wang *et al.* [4] propose a vision research framework, termed parallel vision based on the artificial scenes, computational experiments, and parallel execution (ACP) methodology [5]. ACP methodology is a combination of Artificial scenes, Computational experiments and Parallel execution. They argue that as the partners of the real scenes, the artificial scenes, can be constructed, from which large scale labeled data can be collected for model training and model validation [6]. The core task of the parallel vision framework is to build artificial scenes using basic information extracted from small scale data and human common knowledge. In the artificial scenes, a bundle of computational experiments for different assumptions can be conducted on models before being applied to the real scenes. With a continuous update information collected from the real scenes based on the trained model, the artificial scenes can be updated along with the real scenes, which make it possible to train and validate the models continuously. The process makes an interaction between real scenes and artificial scenes in a parallel style, namely parallel execution. The parallel vision framework offers a long-term and online environment for scene understanding.

The metaverse concept is becoming popular worldwide today. The metaverse allows people to expand their life scenes from real to virtual and make more attempts and mistakes in the virtual scenes. Parallel vision can be viewed as a metaverse approach to understanding the scenes. Meanwhile, to realize adequate knowledge from scenarios for visual reasoning, LeCun [7] proposes constructing the world models for intelligent machines based on the existing information, hoping to encode human knowledge into the world models for reliable and robust scene understanding. From the perspective of world models, the artificial scenes of parallel vision construct a series of world models for scene understanding. With the development of scene understanding, different concepts similar to the parallel vision framework are gradually proposed, which reflects the foresight of the parallel vision framework. However, how to achieve the goal of scene understanding completely based on the parallel vision framework is still being explored.

Li *et al.* [8], [9] conduct the most original and primitive research on artificial scene construction using the game engine Unity3D and conduct computational experiments in artificial scenes for visual tasks [10]. The so-called parallel vision pipeline was finally completed by Li *et al.* [11] after several years, and the results were published in the authoritative Transactions journal. The works by Li *et al.* [8]−[10] make a good start for the exploration of parallel vision, while still contains some shortcomings: 1) The artificial scenes are far from the real scenes, which results in a domain gap for model training for different visual tasks [12]; 2) Even though the labeled data can be collected from the artificial scene directly, while it is time-consuming and laborious for artificial construction manually; 3) Meanwhile, Li *et al.* [11] ignore the real time changes in the real scenes. During the same period, it is not an innovation to train the model on the synthetic datasets [13], [14]. Parallel vision is not only about synthesis or augmentation of images.

In this letter, we review the current research status of parallel vision, and argue that parallel vision is not only works for image synthesis or augmentation but also a framework for total scene understanding by extending the real scenes and interacting the real and artificial scenes in the meantime. In the artificial scenes, we can simulate the real scenes, and model the real scenes in the cyberspace. Meanwhile, we can also predict the feature of the real scenes, with which we can intervene or guide the real scenes in certain aspects, or control the behaviors of the agents. In recent years, various concepts related to scene understanding have advanced greatly, such as metaverse [15], cyber-physical-social systems (CPSS) [16], world models [7] and big models [17]. We combine related items to explain the origin of parallel vision and how to achieve the goal of the parallel vision framework based on existing technologies. Our belief is that the metaverse is an instance of the CPSS, and parallel vision is intended to create a metaverse solution for scene understanding. Meanwhile, the artificial scenes should be organized in a hierarchical style, in which big models can be used as the foundation models containing human knowledge, namely world models, for scene encoding, and the different semantic information is gradually abstracted from bottom to up. Different tasks are learned with explicit targets. Finally, we propose the concrete structure of parallel vision research roadmap for future research.

Corresponding author: Wenwen Zhang.

Citation: W. Zhang, W. Zheng, Q. Li, and F.-Y. Wang, "Parallel vision ⊇ image synthesis/augmentation," *IEEE/CAA J. Autom. Sinica*, vol. 11, no. 3, pp. 782–784, Mar. 2024.

W. Zhang is with the School of Computer Science, Shaanxi Normal University, Xi'an 710119, China (e-mail: 2021136@snnu.edu.cn).

W. Zheng is with the School of Computer Science and Technology, Wuhan 430070, China (e-mail: zwb2022@whut.edu.cn).

Q. Li and F.-Y. Wang are the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: liqiang@qaii.ac.cn; feiyue@ieee.org).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2023.124038

Parallel vision: The parallel vision theory provides a stable execution environment and a theoretical framework that is long-term and controllable for scene understanding. However, it lacks the relevant technology for implementation and support. Parallel vision has been explored by researchers [4], [8], [10] from various viewpoints. Here, we survey on the related works and explain the essence of Parallel Vision from the perspective of the CPSS. In the end, we propose a parallel vision framework that is task-centered and based on the most recent achievements in scene understanding and deep learning.

Parallel vision ≠ computer graphics: Li *et al.* [8], [9] use a 3D modeling engine to model real scenes in 3D space and synthesize images with rendering technology while a big domain gap is existing between the synthetic images and the real images. To adapt the model trained with synthetic images to real scenes, a domain adaptation process [12] must be carried out. Meanwhile, it is time-consuming to construct the realistic 3D models based on 3D game engines. Furthermore, it is impossible to model everything in a 3D game engine manually, which lacks generalization. Labeling real data directly for different visual tasks may be more cost-effective. Anyway Li *et al.* [8], [9] conduct the primitive experiments and make an early exploration in parallel vision. A general pipeline [11] summarizes the technologies for synthesizing images from the 3D modeling game engine and training the model based on generated images. Scale-coordinated 3D scene modeling, realistic textures, and reasonable light changes are necessary for realistic scenes to be rendered. The computer graphics technologies are perfectly suited to meet the needs.

However, parallel vision is not computer graphics or applying computer graphics for artificial scene modeling, but rather to find a method to present the real world in cyberspace. Synthesizing images is just a characteristic of artificial scenes for displaying the presented scenes or for capturing some images for certain goals. The data synthesis process for 3D modeling and 2D images based on computer graphics is not parallel vision. Obviously, using 3D models to build artificial scenes is not a feasible way to achieve the goal of parallel vision for the complex of modeling and managing the huge mount of 3D models.

Parallel vision ≠ synthetic augmentation: Zhang *et al.* [6] propose adding objects of interest to the real scenes for augmented image synthesis based on the real scenes from the perspective of data augmentation. Along with the real scenes, they propose to use changing artificial scenes to simulate the real scenes and complete scene understanding tasks on a long-term basis. The enhanced generation of data enables us to conduct computational experiments to collect more information from real scenes, improving the robustness and generalization of models trained with generated images. The artificial scenes are modeled by combining 3D information and camera angle with the background texture from real scenes to approach the real scenes.

The data augmented artificial scenes lack a representation of the overall scene. It's a challenge to gain a deeper understanding of the scene. The goal of the parallel vision framework is to model real scenes in artificial scenes, in which different information needs to be conducted conveniently, such as 3D geometric information. A more effective representation of the real scenes is still needed, and the details of the scenes need to be parameterized. Parallel vision's needs cannot be met by artificial scenes that use image augmentation.

Parallel vision ≈ CPSS:

1) From CPSS to parallel vision: The 3D modeling method mentioned above, whether it is based on a game engine or data augmentation for artificial scenes modeling, ignores or lacks a clear perception of the real scene. Parallel vision creates artificial scenes that are based on real scenes while expanding them. To gain a better understanding of parallel vision, we start with the fundamental theory of the metaverse, CPSS, to explain what parallel vision is.

Wang [16] first propose the cyber-physical-social systems (CPSS) is to describe the knowledge in CPS that cannot be directly described through analytical models or computational models. To achieve the goal, they propose to create artificial scenes that depict real scenes, which contain human common sense and involve social space. In computer vision, the real scenes are the images and other information, such as 3D geometric information collected by cameras and other depth-related devices. For humans, we model the real scenes in our mental world with only several glimpses and common sense. Parallel vision considers three spaces, and two worlds are involved at the end. Both worlds contain social signals and human common sense. For parallel vision, the key challenge is how to model social signals and human common sense in artificial scenes. Fortunately, the ACP method completes the interaction between physical scenes and artificial scenes, which forms the theoretical framework of parallel vision. Fig. 1 demonstrates the connection between CPSS and the parallel vision framework based on ACP method.

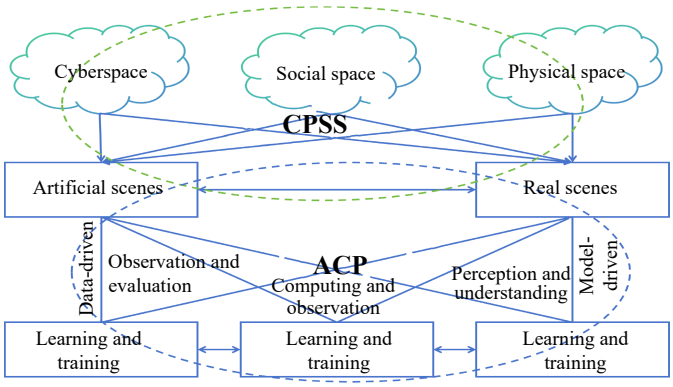

Fig. 1. The connection between CPSS and ACP-based parallel vision.

2) Task-centered scene understanding based on parallel vision: The difference between the visual scenes and other physical scenes is that the representation of the visual scenes (pixel space) and the semantic information (knowledge space) need to be converted, and they are not strictly corresponding. The artificial scenes cannot be modeled directly in 3D geometric space as Li *et al.* [18], [19] have done. In this letter, we propose to use the big model as the fundamental encoding to construct the artificial scenes and use a series of tasks to describe the artificial scenes from bottom to up. Namely, with basic encoding models, all the perception and understanding of the scenes can be expressed as tasks, and the process of learning tasks is to meet the requirements of other tasks for target task. Fig. 2 shows the overall logic diagram of our proposed task-centered scene understanding framework under the parallel vision framework.

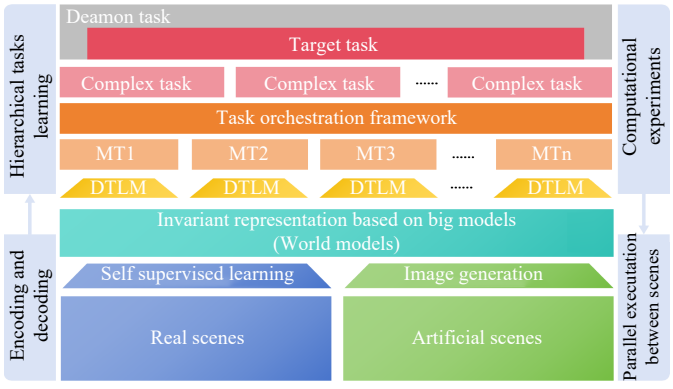

Fig. 2. The structure of parallel vision implementation based on foundation model for scene understanding in a task-centered style: The DTML is short for downstream task learning module, and Mt* means meta tasks.

When human beings want to complete a target task, they divide the target task into different small basic tasks and then assemble them in a certain order, that is, task arrangement, and finally, complete the

target task. The ultimate goal of parallel systems is to describe a dynamic, continuous, and long-term artificial system that perceives and guides real scenes. Our proposal for achieving this goal is to utilize the deamon task for the system's continuous task arrangement. At the same time, the deamon task will continuously obtain the tasks to be completed in the current environment, making the system work continuously for scene understanding.

Conclusion: This letter reviews the current status of the parallel vision framework, and explains it under CPSS. Parallel vision's primary challenge is constructing artificial scenes. As for artificial scenes construction, we should reject directly modeling ANY things in 3D space as Li *et al.* [8], [9] have done. Take the experience from Zhang *et al.* [6] and the development of deep learning for scene understanding community [20], we argue that the parallel vision framework should be constructed based on CPSS with knowledge automation [21], [22] and propose to construct the artificial scenes for parallel vision based on big models for fundamental encoding and a series of hierarchical tasks to understand scenes totally. With encoded scenes based on big models, all actions in scene understanding can be regarded as tasks, and the deamon task is defined to keep the machine functioning at all times, thus realizing a long-term online visual perception framework based on parallel vision.

Acknowledgment: This work was supported by the Natural Science Foundation for Young Scientists in Shaanxi Province of China (2023-JC-QN-0729) and the Fundamental Research Funds for the Central Universities (GK202207008).

References

[1] D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, USA: MIT Press, 1982.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-iid data quagmire of decentralized machine learning," in *Int. Conf. Machine Learning*, 2020, pp. 4387–4398.

[4] K. Wang, C. Gou, N. Zheng, and J. Rehg, "Parallel vision for perception and understanding of complex scenes: Methods, framework, and perspectives," *Artificial Intelligence Review*, vol. 48, pp. 299–329, 2017.

[5] F.-Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Trans. Intelligent Transportation Systems*, vol. 11, no. 3, pp. 630–638, 2010.

[6] W. Zhang, K. Wang, Y. Liu, Y. Lu, and F.-Y. Wang, "A parallel vision approach to scene-specific pedestrian detection," *Neurocomputing*, vol. 394, pp. 114–126, 2020.

[7] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," 2022. [Online], Available: https://openreview.net/forum?id=BZ5a1r-kVsf.

[8] X. Li, K. Wang, Y. Tian, L. Yan, F. Deng, and F.-Y. Wang, "The paralleleye dataset: A large collection of virtual images for traffic vision research," *IEEE Trans. Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2072–2084, 2018.

[9] X. Li, Y. Wang, L. Yan, K. Wang, F. Deng, and F.-Y. Wang, "Paralleleyecs: A new dataset of synthetic images for testing the visual intelligence of intelligent vehicles," *IEEE Trans. Vehicular Technology*, vol. 68, no. 10, pp. 9619–9631, 2019.

[10] Y. Tian, X. Li, K. Wang, and F.-Y. Wang, "Training and testing object detectors with virtual images," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 2, pp. 539–546, 2018.

[11] X. Li, K. Wang, X. Gu, F. Deng, and F.-Y. Wang, "Paralleleye pipeline: An effective method to synthesize images for improving the visual intelligence of intelligent vehicles," *IEEE Trans. Systems*, *Man*, *and Cyber.: Systems*, vol. 53, no. 9, pp. 5545–5556, Sept. 2023.

[12] W. Zhang, J. Wang, Y. Wang, and F.-Y. Wang, "Parauda: Invariant feature learning with auxiliary synthetic samples for unsupervised domain adaptation," *IEEE Trans. Intelligent Transportation Systems*, vol. 23, p. 11, 2022.

[13] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 3234–3243.

[14] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4340–4349.

[15] W. Zheng, L. Yan, W. Zhang, L. Ouyang, and D. Wen, "D→k→i: Data-knowledge-driven group intelligence framework for smart service in education metaverse," *IEEE Trans. Systems*, *Man*, *and Cyber.: Systems*, vol. 53, no. 4, pp. 2056–2061, 2023.

[16] F.-Y. Wang, "The emergence of intelligent enterprises: From CPS to CPSS," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 85–88, 2010.

[17] R. Bommasani, D. A. Hudson, E. Adeli, *et al*., "On the opportunities and risks of foundation models," arXiv preprint arXiv: 2108.07258, 2021.

[18] 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 Intelligent Systems*, vol. 37, no. 4, pp. 18–26, 2022.

[19] X. Li and F.-Y. Wang, "Scenarios engineering: Enabling trustworthy and effective AI for autonomous vehicles," *IEEE Trans. Intelligent Vehicles*, vol. 8, no. 5, pp. 3205–3210, 2023.

[20] W. Zheng, L. Yan, C. Gou, and F.-Y. Wang, "Computational knowledge vision: Paradigmatic knowledge based prescriptive learning and reasoning for perception and vision," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 5917–5952, 2022.

[21] F.-Y. Wang, J. Guo, G. Bu, and J. Zhang, "Mutually trustworthy human-machine knowledge automation and hybrid augmented intelligence: Mechanisms and applications of cognition, management, and control for complex systems," *Frontiers of Infor. Technology & Electronic Engineering*, vol. 23, no. 8, pp. 1142–1157, 2022.

[22] 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. Autom. Sinica*, vol. 9, no. 11, pp. 1899–1908, 2022.