European Journal of AI, Computing & Informatics

Vol. 1 No.1 2025

Review **Open Access**

A Literature Review Study on Knowledge Distillation for Large Models of Image Segmentation

Dongkai Qi [1] and Lim Chia Sien [1,*]

[1] Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia
* Correspondence: Lim Chia Sien, Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia

Received: 16 March 2025
Revised: 20 March 2025
Accepted: 02 April 2025
Published: 05 April 2025

Abstract: With the rapid development of deep learning technology, image segmentation macromodels have achieved remarkable results in many fields. However, these large models often face problems such as high consumption of computational resources and difficulties in deployment. Knowledge distillation, as an effective model compression and optimisation technique, has gradually received attention from researchers. This paper provides a systematic review of the literature related to knowledge distillation for large models of image segmentation, and analyses the current status, advantages and shortcomings of the application of knowledge distillation in large models of image segmentation in terms of improving the AR interactive experience, solving the contradiction between real-time and accuracy, promoting the lightweight and efficient deployment of the model, enhancing the generalization capability of the model, and facilitating the fusion of multimodal data, etc. It also looks forward to the future research direction, aiming to provide a better solution for the research in the related fields. outlook, aiming to provide reference and inspiration for researchers in related fields.

Keywords: image segmentation; knowledge distillation; large models; model compression; AR interaction

1. Introduction

Image segmentation is an important task in the field of computer vision, which divides an image into multiple pixel regions such that each region corresponds to a specific object or part of the image. In recent years, with the rise of deep learning technology, image segmentation macromodels have been widely researched and applied, e.g., they play an important role in the fields of medical image analysis, automatic driving, and intelligent security. However, these large models usually have a large number of parameters and complex structures, resulting in high computational cost, slow inference speed, and difficult to run in real time on resource-constrained devices, which limits the expansion of their practical application scenarios. Knowledge distillation technique is expected to solve this problem by migrating knowledge from complex teacher models to lightweight student models. In this paper, we will review the literature related to knowledge distillation for large models of image segmentation and discuss its applications and research progress in different scenarios.

2. Overview of Knowledge Distillation Techniques

Knowledge distillation is a model compression method, the core idea of which is to use a powerful but complex teacher model to guide the learning of a lightweight student model, so that the student model can inherit most of the performance of the teacher model, while reducing the computational cost and model complexity. In the knowledge distillation process, a loss function is usually defined to measure the difference between the student model and the teacher model, and the student model is trained by optimising this loss function. Common knowledge distillation methods include soft target-based distillation, feature-based distillation, and relationship-based distillation. Each of these methods has its own advantages in different application scenarios, and researchers are constantly exploring and improving knowledge distillation techniques for better application in the optimisation of large models for image segmentation.

3. Research Status of Knowledge Distillation for Image Segmentation Macromodels

3.1. Enhancing AR Interactive Experience

In the field of augmented reality (AR), image segmentation techniques are crucial for achieving the accurate integration of virtual information and real scenes. Knowledge distillation can significantly improve the performance of student models by migrating knowledge from complex teacher models to simple student models, thus enhancing the fusion of virtual information and real scenes [1]. However, the study did not directly mention the enhancement of AR interaction experience. On the other hand, the importance of target detection in mobile AR has been emphasised, as the ability to accurately detect targets in the environment that need to be augmented in real time directly determines the performance of the system. By optimising the target detection algorithm, the interactivity and user experience of the AR system can be significantly improved, but the application of knowledge distillation techniques in it is not explicitly mentioned [2]. A precise surgical navigation system studied the importance of precise navigation for surgical scenarios, with precise image segmentation and visualisation techniques being key to enhancing the AR interaction experience. However, there is a lack of direct discussion on enhancing AR technology's interaction experience in a wider range of scenarios [3]. The BA-KD framework achieves accurate extraction of polyp boundaries through the integration of the boundary segmentation network and the polyp segmentation network, and performs well on multiple datasets. This can significantly improve segmentation accuracy and thus enhance the interactivity and user experience of the AR system, but the study mainly focuses on the field of medical image segmentation, and does not involve other fields [4]. By designing a spatial aggregation module and a channel aggregation module to obtain feature vectors containing global information from both spatial and channel directions, and by enhancing the long-range dependency of the network to improve the ability to express logical anomalies, the accuracy of semantic segmentation was improved, thus indirectly enhancing the interactivity and user experience of the AR system. However, this study mainly focuses on the field of medical image detection and does not cover other fields [5]. The study mainly focuses on industrial anomaly detection and does not directly involve the enhancement of AR interactive experience [6]. Augmented reality technology and personalized recommendation systems with the ability of big models can provide customized suggestions according to passenger habits and situations, and provide navigation information through AR gauges, AR maps, or AR heads-up displays (HUDs). Combined with the learning of user behaviours from the big models, the system can provide personalized navigation, entertainment, and attraction recommendations, demonstrating the potential of big models to enhance the AR interactive experience. However, it cannot directly prove the effect of knowledge distillation on the AR interactive experience [6].

3.2. Solving the Contradiction between Real-Time and Accuracy

In practical applications, image segmentation macromodels often need to make a trade-off between real-time and accuracy. Knowledge distillation can migrate knowledge from complex models to lightweight models without significantly reducing performance, thus achieving high-precision real-time processing on resource-constrained devices, which indirectly illustrates the potential of knowledge distillation in resolving the contradiction between real-time and accuracy. However, it lacks a targeted analysis of AR real-time requirements [7]. Knowledge distillation can effectively improve the classification ability of student networks to achieve network compression, pointing out the potential of knowledge distillation in improving the performance of lightweight models, although the specific application in AR scenarios is not mentioned [8]. Knowledge distillation technology can improve the segmentation accuracy of lightweight models without increasing the computational burden to meet the real-time and high-precision demands of AR, and it has improved the visual inertial navigation system to enhance positioning accuracy and robustness, providing a potential direction for the application of knowledge distillation technology in AR. However, it did not explore in detail the specific implementation method of knowledge distillation technology in solving the conflict between real-time and accuracy, and lacks a detailed understanding of the application of knowledge distillation technology in AR scenarios. The specific implementation method of knowledge distillation technology in solving the contradiction between real-time and accuracy is not explored in detail, and the comparative analysis of different lightweight models is lacking [9]. Applying deep learning methods to target detection in mobile AR is a challenging problem, and there is a contradiction between real-time and accuracy in existing techniques, but knowledge distillation as a solution is not discussed [2]. The use of the DeepSpeed framework and the imperative dataset of synthetic paths of inorganic materials to fine-tune the open-source macromodel Lla-ma2-70b with LoRA fine-tuning technique has provided ideas for solving the problem of the conflict between real-time and accuracy in AR scenarios. However, it is not clearly stated whether it is applicable to lightweight models in AR scenarios [10]. Large models have a large amount of computation and a long inference time, so optimizing the computational efficiency of the model, reducing energy consumption, and speeding up inference is a key topic. This clarifies the real-time problem faced by large models in intelligent networked vehicles, but it cannot directly support the application of knowledge distillation in the semantic segmentation of AR images [6].

Table 1. Solving the contradiction between real-time and accuracy.

No.	Viewpoint	Significance
1	Knowledge distillation can, without significantly reducing performance, transfer knowledge from complex models to lightweight models, thus achieving high-precision real-time processing on resource-constrained devices.	Indirectly explains the potential of knowledge distillation in solving the contradiction between real-time and precision, emphasizing the performance improvement of lightweight models through knowledge distillation.
2	Knowledge distillation can effectively improve the classification ability of student networks, thus achieving network compression.	Points out the potential of knowledge distillation in improving the performance of lightweight models, emphasizing its ability to enhance the precision of lightweight models without increasing computational burden.

3	Knowledge distillation technology can, without increasing computational burden, improve the segmentation accuracy of lightweight models, meeting the real-time and high-precision requirements of AR.	Improves visual inertial navigation systems, enhancing positioning accuracy and robustness, providing a potential direction for the application of knowledge distillation technology in AR.
4	Applying deep learning methods to mobile AR target detection is a challenging issue.	Existing technology has contradictions between real-time and precision.
5	Research utilizes the DeepSpeed framework and inorganic material synthesis path instruction datasets, adopts LoRA fine-tuning technology to fine-tune the open-source large model Llama2-70b, and optimizes the model's hyperparameters, evaluating the optimization effect from two aspects: loss value and model stability during model training.	Provides ideas for solving the contradiction between real-time and precision in AR scenarios.
6	Large models have large computational loads and long inference times, so how to optimize model computational efficiency, reduce energy consumption, and speed up inference is a key issue.	Clarifies the real-time issues faced by large models in intelligent connected vehicles.

3.3. Promoting Model Lightweight and Efficient Deployment

In order to enable image segmentation big models to run efficiently in mobile devices or embedded systems, model lightweighting and efficient deployment are key. Knowledge distillation, as an emerging model compression method, can reduce the number of model parameters and computational cost by migrating the knowledge of complex models into lightweight models, making them more suitable for efficient operation in mobile devices or embedded systems, but it did not mention the specific effect of power consumption reduction [7]. The effective reduction of computation through deep separable convolution and knowledge distillation as a model compression technique can further optimise the lightweight model, reduce the number of parameters and computational cost, and make it more suitable for efficient operation in mobile devices or embedded systems, but it did not address the specific application of knowledge distillation in model lightweighting [2]. Knowledge distillation can migrate knowledge from complex models to lightweight models, reduce the number of model parameters and computational costs, and reduce power consumption while maintaining high accuracy, making it more suitable for efficient operation on resource-constrained devices. However, the study did not elaborate on specific application cases in AR image semantic segmentation scenarios [11]. AR-like terminal devices gather applications in business domains such as display, sensing, interaction, and communication, and can be used as a carrier to show an immersive experience of 3D stereoscopic and virtual-reality fusion, but the study did not mention the role of knowledge distillation technology in model lightweighting and efficient deployment [12,13]. Lightweight and efficient deployment of models through the DeepSpeed framework reduces the memory pressure of a single GPU and provides technical support for the efficient operation of models in mobile devices or embedded systems, but it did not mention whether it is applicable to lightweight models in AR scenarios [10]. The importance of hardware and algorithm synergy, such as developing more efficient sparse matrix operations and low-precision computation support hardware to improve the deployment efficiency of pruning and quantisation techniques, is repeatedly mentioned, but

the role of knowledge distillation techniques in model lightweighting is not explicitly discussed [6].

Table 2. Solving the contradiction between real-time and accuracy.

No.	Viewpoint	Significance
1	Knowledge distillation, as an emerging model compression method, can transfer knowledge from complex models to lightweight models, reducing model parameters and computational costs, making it more suitable for efficient operation on mobile devices or embedded systems.	Points out the application of knowledge distillation in model light-weighting, emphasizing its optimization of model parameters and computational costs.
2	Through deep separable convolution to effectively reduce computational volume.	Knowledge distillation, as a model compression technology, can further optimize lightweight models, reduce parameters and computational costs, making it more suitable for efficient operation on mobile devices or embedded systems.
3	Knowledge distillation can transfer the knowledge of complex models to lightweight models, reducing model parameter volume and computational costs.	Clearly points out the advantages of knowledge distillation in model lightweighting and efficient deployment, allowing it to maintain high precision while reducing model parameter volume and computational costs, making it more suitable for efficient operation on resource-constrained devices, reducing power consumption.
4	AR terminal devices integrate display, sensing, interaction, communication, and other business applications, providing an immersive experience that combines three-dimensional reality and virtual reality.	Emphasizes the advantages of AR devices as large model carriers, clarifying the importance of AR devices in large model applications.
5	By using the DeepSpeed framework, model light-weighting and efficient deployment are achieved, reducing the memory pressure of a single GPU. This provides technical support for efficient operation of models on mobile devices or embedded systems.	Optimizes the deployment efficiency of models.
6	Hardware and algorithm collaboration: develop more efficient sparse matrix computation and low-precision computing support hardware to enhance the deployment efficiency of pruning and quantization technologies.	Emphasizes the importance of hardware and algorithm collaboration optimization.

3.4. Enhance Model Generalisation Ability

The generalisation ability of large models for image segmentation is crucial when facing complex and changing scenes. Knowledge distillation can significantly improve the generalisation ability of student models in complex scenarios by transferring the "dark knowledge" of the teacher's model, so as to maintain stable performance in changing environments, but it lacks a specific analysis of AR scenarios [7]. Model lightweighting itself may help improve generalisation ability and provide support for efficient model operation, but there is insufficient direct evidence and in-depth analysis of generalisation ability improvement [8]. AR scenarios are complex and diverse, and knowledge distillation can enhance the adaptability of models to different scenarios, enhance their generalisation ability, and enable them to maintain stable performance in changing environments [14], but it does not explore in detail the specific implementation methods of knowledge distillation technology in enhancing the generalisation ability of models, and it lacks a comparative analysis of different scenarios [9]. It is difficult to apply deep learning target detection in mobile AR due to the limited computing power of mobile AR devices, but the role of knowledge distillation in enhancing model generalisation ability was not discussed [2]. The BA-KD model performs well on multiple datasets, not only in the colon polyp segmentation task, but also in the dermatoscopy image segmentation dataset, which demonstrates a strong generalisation ability, and improves the model's adaptability to different scenarios and enhances its generalisation ability through knowledge distillation [15]. However, the study did not address the complex diversity of AR scenes [4]. By designing a selective fusion module (SFM) to amplify features containing important information and enhance the model's understanding of global information, the model's adaptability to different types of anomalies is improved [16]. However, the study was mainly aimed at industrial anomaly detection and did not deal with the complexity and diversity of AR scenarios [5]. The seamless integration of vision, reality, machine, and virtual space can be achieved through AR technology, which will ultimately drive a new round of industrial change, but the impact of knowledge distillation technology on the generalisation ability of the model was not specifically addressed [12]. Expanding the model parameter scale can realise the "emergence ability" of the Smart Driver model to a certain extent, emphasising the impact of the model scale on the generalisation ability, but it does not involve the application of knowledge distillation technology to directly support its application in the semantic segmentation of AR images [6].

3.5. Facilitating Multimodal Data Fusion

Multimodal data fusion is of great significance in image segmentation tasks, as it can make full use of different types of data information and improve segmentation accuracy. Knowledge distillation can take advantage of the features of data such as unlabelled and cross-modal, which has a significant enhancement effect on model enhancement and provides a new solution for multimodal data fusion, but it lacks a specific discussion on multimodal data in AR scenarios [7]. The intrinsic association of multi-scale unimodal multimodal potential semantic information in multimodal medical images was revealed, and the intrinsic association of different modal information was explored through multimodal image analysis, which provided a theoretical basis for multimodal data fusion [17]. However, it could not directly prove the contribution of multimodal data fusion to knowledge distillation technology [3]. Attention graph features are adopted as the intermediate layer knowledge to make up for the lack of single knowledge information in the output layer, enriching the feature knowledge by combining two kinds of knowledge distillation to ensure the diversity of knowledge information in the teacher network model. This provides a new idea for multimodal data fusion in AR scenes, but it does not explicitly explore the specific application of multimodal data fusion, and the practical effect of multimodal data fusion in semantic segmentation of AR images needs further research and validation [11].

The multimodal large model can handle multiple types of data inputs such as vision, language, radar, and LIDAR, which is suitable for processing complex environmental information, and through multimodal data fusion, accurate perception and decision support can be generated, clarifying the necessity of multimodal data fusion, but it does not address the application of knowledge distillation technology in AR image semantic segmentation [13].

4. Research Summary and Outlook

This paper provides a systematic review of the literature related to knowledge distillation for large models of image segmentation, and analyses the current status and research progress of the application of knowledge distillation technology in large models of image segmentation from multiple perspectives. In general, knowledge distillation technology shows great potential in improving AR interactive experience, solving the contradiction between real-time and accuracy, promoting model lightweight and efficient deployment, enhancing model generalisation capability, and facilitating multimodal data fusion, providing strong support for the optimisation and application expansion of image segmentation macromodels. However, there are still some shortcomings in the current research, such as fewer specific application cases in AR scenarios, lack of in-depth comparative analyses on the implementation methods and effects of knowledge distillation techniques in different scenarios, and some studies fail to fully combine the practical application requirements for targeted exploration.

Future research can be explored in depth in the following directions: first, to further strengthen the research on the application of knowledge distillation technology in AR scenarios, combining with the actual AR application requirements, exploring the optimisation strategies and specific implementation methods of knowledge distillation technology in enhancing AR interactive experience, real-time, accuracy, etc., so as to provide stronger technical support for the development of AR technology; second, to further study the specific mechanisms and optimisation methods of knowledge distillation technology The second is to thoroughly study the specific mechanism and optimization method of knowledge distillation technology in solving the contradiction between real-time and accuracy, and find a more suitable optimization scheme for image segmentation large model by comparing and analysing different lightweight models and knowledge distillation methods, in order to better satisfy the real-time and accuracy requirements of practical applications; the third is to combine the idea of collaborative optimization of hardware and algorithms, and explore the combination of knowledge distillation technology with hardware acceleration, model pruning, quantization and other technologies to The combination of knowledge distillation technology with hardware acceleration, model pruning, quantisation and other techniques can further promote the lightweight and efficient deployment of large models for image segmentation, reduce the computational cost and power consumption of the models, and enable them to run efficiently on more devices; fourth, we will strengthen the in-depth study of knowledge distillation technology in enhancing the generalisation ability of the models and the fusion of multimodal data, and explore how to better mine and make use of the knowledge information in the models through the knowledge distillation technology to increase the generalisation ability and the fusion of multimodal data of the models in the complex and variable scenarios. The generalisation ability of the model and the fusion ability of multimodal data are improved to provide theoretical and technical support for the application of large models of image segmentation in a wider range of fields; Fifthly, we pay attention to the research on the interpretability of knowledge distillation technology, and through the in-depth analysis of the mechanism of knowledge transfer and model learning in the process of knowledge distillation, we can improve the interpretability and reliability of the knowledge distillation technology, which will provide help for the researcher to understand and apply the knowledge distillation technology in a better way.

References

1. Y. Feng, X. Sun, W. Diao, J. Li and X. Gao, "Double Similarity Distillation for Semantic Image Segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 5363-5376, 2021, doi: 10.1109/TIP.2021.3083113.
2. S. J. Cheng, Q. X. Zhao, X. Y. Zhang, N. Yadikar, and K. Ubul, "A review of knowledge distillation in object detection," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3288692.
3. D. Qin *et al.*, "Efficient medical image segmentation based on knowledge distillation," *IEEE Trans. Med. Imaging*, vol. 40, no. 12, pp. 3820-3831, Dec. 2021, doi: 10.1109/TMI.2021.3098703.
4. Q. Dou, Q. Liu, P. A. Heng, and B. Glocker, "Unpaired multi-modal segmentation via knowledge distillation," *IEEE Trans. Med. Imaging*, vol. 39, no. 7, pp. 2415-2425, July 2020, doi: 10.1109/TMI.2019.2963882.
5. S. Deng et al., "LHAR: Lightweight human activity recognition on knowledge distillation," *IEEE J. Biomed. Health Informatics*, vol. 28, no. 11, pp. 6318-6328, Nov. 2024, doi: 10.1109/JBHI.2023.3298932.
6. Z. Li et al., "When object detection meets knowledge distillation: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10555-10579, Aug. 2023, doi: 10.1109/TPAMI.2023.3257546.
7. W. Zhou, H. Zhang, and W. Qiu, "Differential modal multistage adaptive fusion networks via knowledge distillation for RGB-D mirror segmentation," *IEEE Trans. Big Data*, doi: 10.1109/TBDATA.2024.3505057.
8. Z. Tu, X. Liu, and X. Xiao, "A general dynamic knowledge distillation method for visual analytics," *IEEE Trans. Image Process.*, vol. 31, pp. 6517-6531, 2022, doi: 10.1109/TIP.2022.3212905.
9. P. Taveekitworachai, P. Suntichaikul, C. Nukoolkit, and R. Thawonmas, "Speed up! Cost-effective large language model for ADAS via knowledge distillation," in *2024 IEEE Intell. Vehicles Symp. (IV)*, Jeju Island, Korea, Republic of, 2024, pp. 1933-1938, doi: 10.1109/IV55156.2024.10588799.
10. C. Chen, Q. Dou, Y. Jin, Q. Liu, and P. A. Heng, "Learning with privileged multimodal knowledge for unimodal segmentation," *IEEE Trans. Med. Imaging*, vol. 41, no. 3, pp. 621-632, Mar. 2022, doi: 10.1109/TMI.2021.3119385.
11. P. Zhao, Y. Hou, Z. Yan, and S. Huo, "Text-driven medical image segmentation with text-free inference via knowledge distillation," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1-15, 2025, Art no. 5011715, doi: 10.1109/TIM.2025.3545506.
12. F. Xiong, C. Shen, and X. Wang, "Generalized knowledge distillation for unimodal glioma segmentation from multimodal models," *Electronics*, vol. 12, no. 7, p. 1516, 2023, doi: 10.3390/electronics12071516.
13. T. Wan, N. Canagarajah, and A. Achim, "Segmentation-driven image fusion based on alpha-stable modeling of wavelet coefficients," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 624-633, Jun. 2009, doi: 10.1109/TMM.2009.2017640.
14. D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proc. IEEE*, vol. 103, no. 9, pp. 1449-1477, Sept. 2015, doi: 10.1109/JPROC.2015.2460697.
15. W. Chen, G. Yang, and D. Qi, "Comprehensive evaluation of college students' physical health and sports mode recommendation model based on decision tree classification model," *Comput. Intell. Neurosci.*, vol. 2022, pp. 5504850, Jul. 2022, doi: 10.1155/2022/5504850.
16. W. Chen, S. A. S. K. B. Syed Ali, H. Zulnaidi, and D. Qi, "Research on intelligent analysis of healthy training progress of teenage sports athletes using various modalities," *Sustainability*, vol. 14, no. 24, Art no. 16556, 2022, doi: 10.3390/su142416556.
17. H. Huang and D. Qi, "Application of improved CNN technology in medical imaging course," *Higher Educ. Orient. Stud.*, vol. 2, no. 6, pp. 40-49, doi: 10.54435/heos.v2i6.85.