Cross-Modal Graph Learning for Perivascular Spaces Segmentation

Tao Chen[1,†], Dan Zhang[2,†], Xi Long[1], Marcel Breeuwer[1], Sveta Zinger[1], Peiyu Huang[3,✉], and Jiong Zhang[4,✉]

[1] Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands
[2] School of Cyber Science and Engineering, Ningbo University of Technology, Ningbo, China
[3] Department of Radiology, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
[4] Laboratory of Advanced Theranostic Materials and Technology, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
{huangpy@zju.edu.cn;jiong.zhang@ieee.org}

Abstract. Perivascular spaces (PVS), also known as Virchow-Robin spaces, are critical biomarkers for diagnosing cerebral small vessel disease (CSVD). Quantifying PVS visible in magnetic resonance imaging (MRI) is essential for understanding their relationship with various neurological disorders. Traditional methods for assessing PVS rely on visual scoring of MRI images, which is time-consuming, subjective, and unsuitable for large-scale studies. Additionally, due to their small size, scattered distribution, and complex morphology, PVS can easily be confused with neighboring structures, posing significant challenges for their accurate extraction. In this paper, we propose a novel graph interaction-enhanced model based on vision-language modeling (VLM) technology for accurate PVS extraction from MRI. Our approach leverages textual information to guide image feature extraction and employs a graph structure to enhance cross-modal interactions, facilitating the reasoning of relationships between different modalities. Furthermore, we introduce a cross-modal attention mechanism for global feature alignment and an attention-based dynamic fusion module to effectively integrate multi-modal information, improving the accuracy of PVS segmentation. Validated on an independent T1-weighted dataset, our model demonstrates superior performance in capturing both global and local information, addressing the limitations of traditional image-only models and providing a robust solution for PVS segmentation in complex clinical scenarios.

Keywords: PVS · MRI · GCN · VLM · Cross Attention

[†] These authors contributed equally to this work.

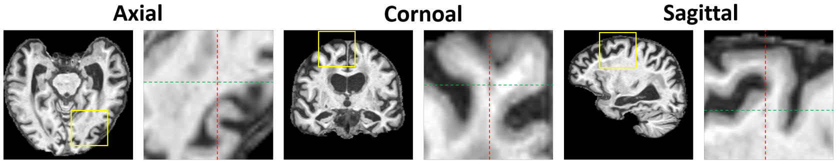

Fig. 1. T1-weighted MR images of perivascular spaces (PVS) are displayed in axial, coronal, and sagittal planes. The highlighted areas in the yellow boxes are magnified on the right side for detailed visualization.

1 Introduction

The space around small intracranial blood vessels is known as the perivascular space (PVS) or Virchow-Robin space [1]. Abnormal manifestations of the PVS, such as enlargement or an increase in the number of PVS, have been shown to be associated with various neurological disorders [2], including small-vessel disease [3], Alzheimer's disease [4], and multiple sclerosis [5]. These associations make the PVS an important biomarker for studying brain health and diseases[6]. Traditionally, clinicians have relied on visual scoring based on 3T MRI images to assess PVS [7, 8], but this approach requires extensive clinical experience. It is not only time-consuming but also subject to limitations such as lower and upper boundary effects, making it unsuitable for large-scale studies and clinical applications. Additionally, PVS are usually small, scattered, and morphologically complex, and depending on their orientation and angle, they can exhibit linear and circular shapes, among others, and can be easily confused with neighbouring structures [9], as shown in Figure 1. Therefore, there is an urgent need for automated methods to address the challenges of PVS segmentation.

Previously, researchers developed many segmentation methods for PVS in anticipation of accurate quantification and analysis of the volume of PVS. These methods can be classified into traditional methods based on information such as low-level features (e.g., grey values, morphological features) [10–12] and a priori knowledge of the image, and data-driven deep learning-based methods [13–15]. Among these methods, Ballerini [10] et al. achieved PVS segmentation by optimizing and evaluating the parameters of the Frangi filter using an ordered logit model and a visual rating scale as substitutes for ground truth labels. Niazi [11] et al. and Cai [12] et al. employed edge detection and K-means clustering, respectively, for PVS segmentation. While interpretable, these methods require manual tuning and are time-consuming, often failing to find globally optimal parameters. With the advancement of deep learning technology, many deep learning-based techniques have been proposed to improve PVS segmentation. Huang [14] developed a nnUNet-based method suitable for multi-center clinical imaging; however, it still struggles to handle small and dispersed PVS structures. Lian [15] designed a multi-scale CNN model that provides rich contextual information, but this method depends on 7T MRI data, whereas most clinical imaging uses 3T MRI, making it less practical for real-world applications.

Therefore, despite these advancements, there remains an urgent need for flexible and accurate methods capable of segmenting small, scattered, and low-contrast PVS structures in complex clinical scenarios.

Recently, Vision-Language Modeling (VLM) [16, 17] has gained attention for learning joint image-text representations and demonstrating strong generalization across vision tasks. In PVS segmentation, where structures are often small, scattered, and low-contrast, traditional image-only models struggle to capture morphological details. For this reason, we introduce the VLM framework to address this, leveraging linguistic information to better understand PVS morphology and distinguish it from similar structures. Textual descriptions also provide spatial cues, aiding precise PVS localization and reducing interference from surrounding anatomy. Meanwhile, to further enhance fine-grained detail capture, we integrate VLM with a Graph Convolutional Network (GCN) [18–20]. GCN models interactions between entities through graph structures, enabling effective cross-modal interaction and better capturing complex local relationships, thereby improving the representation of subtle PVS features. Finally, we propose a cross-modal cross-attention module and dynamic multi-modal fusion module for precise global feature alignment and modality integration. This approach retains VLM's cross-modal strengths while significantly boosting segmentation accuracy for challenging PVS cases, overcoming the limitations of image-only models. The main contributions are summarized as follows:

(a) We propose a vision-language model for PVS segmentation, leveraging textual information to enhance the model's understanding of PVS morphology. This innovation addresses the limitations of image-only methods in identifying small, low-contrast PVS structures, improving segmentation accuracy.
(b) We design a novel cross-modal graph interaction enhancement module that models relationships between different modalities through a graph structure. By utilizing textual cues to capture fine-grained local features, it enables precise cross-modal feature interaction, enhancing the model's ability to recognize PVS boundaries and morphology.
(c) We propose a cross-modal attention strategy and a dynamic multimodal fusion module to achieve global feature alignment and seamless information integration. This approach enhances multimodal complementarity and optimizes segmentation performance.

2 Proposed Method

The proposed PVS segmentation framework consists of four main components: a language-vision framework, a multimodal graph interaction enhancement module, a cross-modal global information alignment module, and a multimodal fusion module, as shown in Fig. 2. Given an image and a natural language description of the target object, the model generates pixel-level segmentation predictions. In the language-vision framework, BERT, a pre-trained language model developed by Google, is used as the text encoder. The image encoder is designed as a nested U-network for richer feature learning.

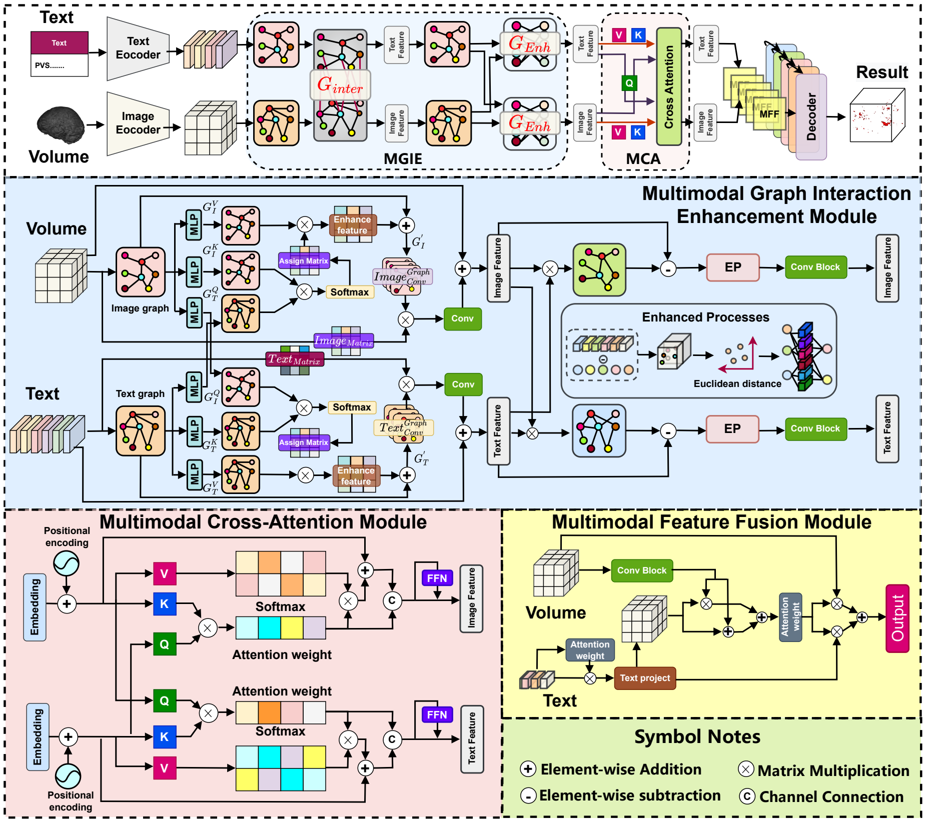

Fig. 2. Schematic diagram of the proposed PVS segmentation network, containing the overall framework of the network and specific details of the proposed modules.

2.1 Multimodal Graph Interaction Enhancement Module

Multimodal Graph Interaction: For the VLMs framework, better feature alignment as well as inter-modal information interaction is the key to improve the model performance. The MGIE module first performs feature interaction between text and image through the graph structure, and then subsequently performs complementary enhancement using the textual and image information respectively. The specific steps are as follows, given the input features $F_{\mathrm{I}} \in \mathbb{R}^{D\times H\times W\times C}$ and $F_{\mathrm{T}} \in \mathbb{R}^{T\times C}$, where I denotes the image and T denotes the text. The features are transformed into embedded representations $G_{\mathrm{T}} \in \mathbb{R}^{K\times C}$ and $G_{\mathrm{I}} \in \mathbb{R}^{K\times C}$ of graph nodes through graph projection G^{pro} operations. We parameterise G using $L \in \mathbb{R}^{C\times K}$ and $\sigma \in \mathbb{R}^{C\times K}$, where is L randomly initialised with $K = 10$ center notes and σ is a measure of the range of the input feature in relation to that node. The nodes are computed as follows:

$$S_k^i = \frac{\exp(-\|(f_i - l_k)/\sigma_k\|_2^2/2)}{\sum_{p=1}^{K} \exp(-\|(f_i - l_p)/\sigma_p\|_2^2/2)}, i \in (D \times H \times W) \text{ or } T \tag{1}$$

$$node_k = \frac{node_k^*}{\|node_k^*\|_2}, node_k^* = \frac{\sum_{i=1}^{I} S_k^i (f_i - l_k)/\sigma_k}{\sum_{i=1}^{I} S_k^i}. \tag{2}$$

Where S is the soft assignment matrix, obtained from the residuals of each input feature with respect to the centre node and normalised by the standard deviation, which indicates the strength of the input features with respect to the node, and then the features are aggregated to the node to generate the final node representation. In addition, the edges of the graph are implicitly represented in the assignment matrix as the relationship between the input features and the nodes. Thus the adjacency matrix of the graph can be expressed as: $A_{\text{dj}} = (G^{\text{T}} \times G)$. Next the graph interaction G^{inter} operation models the relationships between graphs. As shown in Fig. 2, G^{inter} adopts the cross-attention mechanism, G_{T} and G_{I} are first converted to Q_{graph}, K_{graph} and V_{graph} by MLP, and then the matrix multiplication method is used to construct the attention score matrix $A_{\text{dj}}^{\text{T-I}} = (Q_{\text{T}}^{\text{T}} \times K_{\text{I}})$ and $A_{\text{dj}}^{\text{I-T}} = (Q_{\text{I}}^{\text{T}} \times K_{\text{T}})$, and weighted to the original V_{graph} to get the final output. $\mathcal{W}$ denotes the weighting parameter to adjust the importance of G^{inter} relative to $G_{\text{T}}^{'}$ and $G_{\text{I}}^{'}$.

$$G_{\text{T}}^{'} = \mathcal{W}(A_{\text{dj}}^{\text{T-I}} \times V_{\text{T}}) + G_{\text{T}}, G_{\text{I}}^{'} = \mathcal{W}(A_{\text{dj}}^{\text{I-T}} \times V_{\text{I}}) + G_{\text{I}} \tag{3}$$

After executing G^{inter}, the graph reasoning G^{rea} is utilized to send $G_{\text{T}}^{'}$ and $G_{\text{I}}^{'}$ as inputs to the constructed graph convolution $\mathcal{G}$ for inference updating, and the interacted features are output using the reprojection operation.

$$G_{\text{T}}^{''} = G^{\text{rea}}\{\mathcal{G}_{\text{T}}(A_{\text{dj}}^{\text{T}} \times G_{\text{T}}^{'} \times W_{\text{T}})\}, F_{\text{T}}^{'} = G^{\text{repro}}\{M_{\text{T}} G_{\text{T}}^{''\text{T}} + F_{\text{T}}\} \tag{4}$$

$$G_{\text{I}}^{''} = G^{\text{rea}}\{\mathcal{G}_{\text{I}}(A_{\text{dj}}^{\text{I}} \times G_{\text{I}}^{'} \times W_{\text{I}})\}, F_{\text{T}}^{'} = G^{\text{repro}}\{M_{\text{I}} G_{\text{I}}^{''\text{I}} + F_{\text{I}}\} \tag{5}$$

Where M is the assignment matrix in G^{pro}, used to reconstruct the graph features into the original features, and W denotes the learnable weight parameters. **Multimodal Graph Enhancement:** The post-interaction text and image features not only enhance the representation of intra-modal information, but also incorporate cross-modal interaction information. Therefore, in the graph enhancement step, we first weight the post-interaction text features and image features with $F_{\text{fusion}} = F_{\text{T}}^{'} \times F_{\text{I}}^{'}$ using the channel multiplication operation, and then transform them into text-based node embeddings $T = \{t_1, t_2 \ldots, t_k\}$ and image-based node embeddings $I = \{i_1, i_2 \ldots, i_k\}$ via G^{pro}, respectively. Subsequently, the edge relationship between the two is constructed by means of K-NN, which in turn connects $f_i^{\text{T}} \in F_{\text{fusion}}^{\text{T}}$ and t_j as well as $f_i^{\text{I}} \in F_{\text{fusion}}^{\text{I}}$ and i_j are connected by means of K-NN, which in turn generates the text-support graph G_{I}^{T} and image-support graph G_{T}^{I}. Formally, the text-edge embedding and the image-edge embedding can be expressed as:

$$T_{i,j}^{\prime} = f_{\text{Conv}}(f_i^{\text{T}} - t_j), \;\; I_{i,j}^{\prime} = f_{\text{Conv}}(f_i^{\text{I}} - t_j), \tag{6}$$

where f_{Conv} is used to compute the feature difference embedding. In addition, the number of nearest neighbour edges is set to 5 for calculating the edge relationships. Finally maximum pooling aggregation is performed on all the support nodes connected to f_i^{T} and f_i^{I} to get the most significant features.

2.2 Multimodal Cross-Attention Module

The graph structure can effectively model local relationships and topological structures but has limitations in capturing global relationships. Therefore, we design a multimodal cross-attention module to model global relationships and non-local interactions, further refining the interaction-enhanced features. Specifically, we transform the image and text features into Key, Value, and Query representations, then compute attention weights by interacting Query with Key and apply these weights to Value. Finally, residual connections, feature concatenation, and a feed-forward network are employed to fuse and optimize the interaction features, producing enhanced image and text representations.

2.3 Multimodal Feature Fusion Module

The final step fuses the two modalities for output. Traditional methods use pixel-wise concatenation or feature multiplication, but expanding text features to match image dimensions weakens positional information. To address this, we design a dynamic fusion module using an attention mechanism. First, self-attention-based weighted pooling highlights relevant text features while reducing redundancy. Then, additive and multiplicative interactions between text and image features are fed into an attention module to generate dynamic spatial weights, representing the importance of each feature at different locations. Finally, these weights refine both modalities to produce the fused features.

3 Experimental Results

Dataset: The method was validated on a private dataset consisting of 3D T1-weighted images from 50 subjects. The annotations were performed manually by two experienced clinicians and subsequently reviewed and refined by a senior expert. The acquisition of all images was approved by the relevant regulatory authorities, and informed consent was obtained from all patients, in compliance with the Declaration of Helsinki.
Implementation Details: Our method was implemented using the PyTorch framework and trained on two NVIDIA GeForce GTX 4090 GPUs. The model was optimized using the Adam optimizer with a cosine annealing learning rate decay strategy. The initial learning rate was set to 0.001, with a batch size of 4, and training was conducted for 1,000 epochs. During training, the original images were cropped to a size of 96×96×96 and subjected to random rotations along three axes. Finally, 5-fold cross-validation was employed to evaluate the performance of the model.

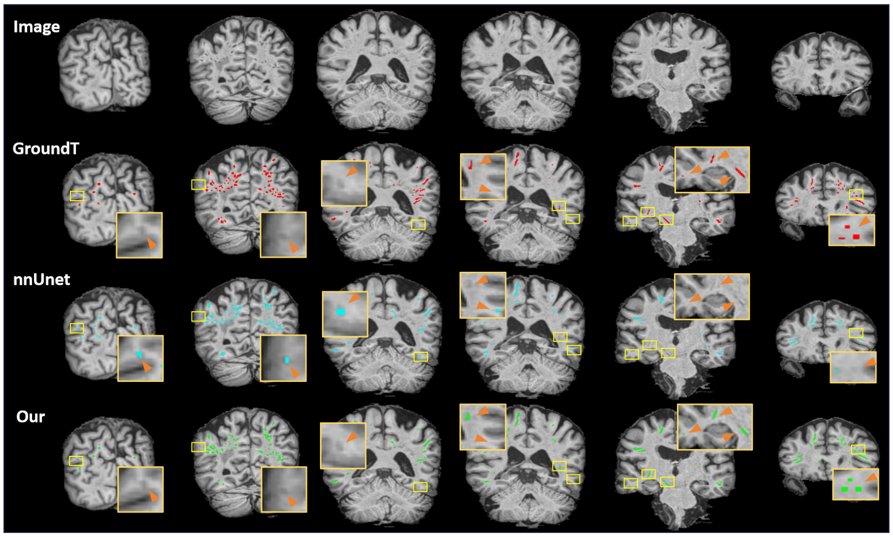

Fig. 3. Comparison of segmentation results with nnUnet [21], the yellow box shows the comparison of segmentation details with square large processing.

Table 1. Performance comparisons for PVS segmentation.

Methods	PVS segmentation			
	DICE	IoU	Recall	Precision
CS2-Net [22]	0.3530	0.2226	0.3138	0.4908
UNetr [23]	0.3809	0.2565	0.3625	0.5017
Swin-UNetr [24]	0.4230	0.2728	0.4042	0.5128
SHIVSV1-Net [13]	0.4966	0.3109	0.4357	0.5344
SHIVSV2[13]	0.5012	0.3457	0.4574	0.5417
PINGU [25]	0.5733	0.4079	0.5129	0.6034
nnUNet [21]	0.5797	0.4127	0.5580	0.6107
Proposed	**0.6042**	**0.4390**	**0.5771**	**0.6571**

Comparison with State-Of-The-Arts: To evaluate the performance of our model, we compared it with several classical 3D medical image segmentation methods, including the CS2-Net [22], UNETR [23], SwinUNet [24], and nnU-Net [21]. Additionally, we evaluated our model against publicly available T1-weighted image PVS segmentation methods, namely PINGU [25], SHIVSV1 and V2 [13]. PINGU [25] is built upon nnU-Net [21], while SHIVSV1 and V2 [13] are based on ResNet3D and 3D U-Net, respectively. For a fair comparison, we utilized the pre-trained weights provided by these methods for validation.

Table 2. Ablation results for PVS segmentation.

Methods	PVS segmentation			
	DICE	IoU	Recall	Precision
Backbone	0.5338	0.3641	0.4539	0.5361
Backbone + VLM	0.5515	0.3857	0.4716	0.6729
Backbone + VLM + MGIE	0.5828	0.4158	0.5510	0.6270
Backbone + VLM + MGIE + MCA	0.5935	0.4260	0.5613	0.6515
Backbone + VLM + MGIE + MCA + MFF	**0.6042**	**0.4390**	**0.5771**	**0.6571**

The quantitative results are shown in Table 1. We selected the Dice similarity coefficient (Dice) and Intersection over Union (IoU) as the primary evaluation metrics, with Precision used to assess false positives and Recall to highlight false negatives. The comparison results demonstrate that the proposed method achieves the best performance, with a Dice score of 60%. Fig. 3 presents the visualization results. Notably, to more comprehensively compare the performance of different methods in various regions of the image, we conducted a visual comparison with nnU-Net, which achieved the best performance among the baseline methods. As shown in the figure, our method achieves outstanding segmentation across different PVS regions, whereas nnU-Net exhibits several false positive and false negative segmentations.

We attribute these results to the use of text guidance for image segmentation, which enables the model to better locate target regions while eliminating interference from non-target areas. Additionally, improved modality interaction and alignment further enhance the accuracy of the VLM-based model.

Ablation Study: To verify the effectiveness of the proposed framework and its individual modules for PVS segmentation, we conducted an ablation study based on a nested U-shaped CNN framework as the backbone. We successively added the VLM structure, the multimodal graph interaction enhancement module, the cross-modal attention module, and the dynamic fusion module to analyze their impact. Table 2 presents the quantitative results of this analysis. The results show that VLM significantly improves the baseline performance of PVS segmentation. Furthermore, with the addition of the local graph interaction enhancement module, the global cross-modal attention module, and the multimodal fusion module, segmentation accuracy continuously improves. Ultimately, our method achieves a Dice score of 60%.

4 Conclusion

In summary, this study proposes a novel approach to address the challenge of PVS segmentation in T1-weighted MRI images. The proposed method adopts a VLM framework, integrating graph structures to enhance local cross-modal information interaction. Additionally, cross-attention is leveraged for global cross-modal learning, followed by a dynamic fusion strategy for modality integration. The results demonstrate that utilizing text to provide prior guidance for image segmentation, along with sufficient cross-modal interaction and fusion, can significantly improve PVS segmentation accuracy. This has important implications for aiding the clinical diagnosis of various neurological diseases. In the future, we will explore the application of this method in multimodal imaging.

Acknowledgment This work was supported in part by the National Science Foundation Program of China (62371442), and in part by the Zhejiang Provincial Natural Science Foundation (LQ23F010002, LZ23F010002, LR24F010002).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zhang, E., Inman, C., Weller, R.: Interrelationships of the pia mater and the perivascular (virchow-robin) spaces in the human cerebrum. Journal of anatomy **170** (1990) 111
2. Charidimou, A., Boulouis, G., Pasi, M., Auriel, E., van Etten, E.S., Haley, K., Ayres, A., Schwab, K.M., Martinez-Ramirez, S., Goldstein, J.N., et al.: Mri-visible perivascular spaces in cerebral amyloid angiopathy and hypertensive arteriopathy. Neurology **88**(12) (2017) 1157–1164
3. Francis, F., Ballerini, L., Wardlaw, J.M.: Perivascular spaces and their associations with risk factors, clinical disorders and neuroimaging features: a systematic review and meta-analysis. International Journal of Stroke **14**(4) (2019) 359–371
4. Zeng, Q., Li, K., Luo, X., Wang, S., Xu, X., Jiaerken, Y., Liu, X., Hong, L., Hong, H., Li, Z., et al.: The association of enlarged perivascular space with microglia-related inflammation and alzheimer's pathology in cognitively normal elderly. Neurobiology of Disease **170** (2022) 105755
5. Miyata, M., Kakeda, S., Iwata, S., Nakayamada, S., Ide, S., Watanabe, K., Moriya, J., Tanaka, Y., Korogi, Y.: Enlarged perivascular spaces are associated with the disease activity in systemic lupus erythematosus. Scientific reports **7**(1) (2017) 12566
6. Duering, M., Biessels, G.J., Brodtmann, A., Chen, C., Cordonnier, C., de Leeuw, F.E., Debette, S., Frayne, R., Jouvent, E., Rost, N.S., et al.: Neuroimaging standards for research into small vessel disease—advances since 2013. The Lancet Neurology **22**(7) (2023) 602–618
7. Doubal, F.N., MacLullich, A.M., Ferguson, K.J., Dennis, M.S., Wardlaw, J.M.: Enlarged perivascular spaces on mri are a feature of cerebral small vessel disease. Stroke **41**(3) (2010) 450–454
8. Zhu, Y.C., Tzourio, C., Soumaré, A., Mazoyer, B., Dufouil, C., Chabriat, H.: Severity of dilated virchow-robin spaces is associated with age, blood pressure, and mri markers of small vessel disease: a population-based study. Stroke **41**(11) (2010) 2483–2490

9. Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., T O'Brien, J., Barkhof, F., Benavente, O.R., et al.: Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology **12**(8) (2013) 822–838
10. Ballerini, L., Lovreglio, R., Valdés Hernández, M.d.C., Ramirez, J., MacIntosh, B.J., Black, S.E., Wardlaw, J.M.: Perivascular spaces segmentation in brain mri using optimal 3d filtering. Scientific reports **8**(1) (2018) 2132
11. Niazi, M., Karaman, M., Das, S., Zhou, X.J., Yushkevich, P., Cai, K.: Quantitative mri of perivascular spaces at 3t for early diagnosis of mild cognitive impairment. American Journal of Neuroradiology **39**(9) (2018) 1622–1628
12. Cai, K., Tain, R., Das, S., Damen, F.C., Sui, Y., Valyi-Nagy, T., Elliott, M.A., Zhou, X.J.: The feasibility of quantitative mri of perivascular spaces at 7 t. Journal of neuroscience methods **256** (2015) 151–156
13. Boutinaud, P., Tsuchida, A., Laurent, A., Adonias, F., Hanifehlou, Z., Nozais, V., Verrecchia, V., Lampe, L., Zhang, J., Zhu, Y.C., et al.: 3d segmentation of perivascular spaces on t1-weighted 3 tesla mr images with a convolutional autoencoder and a u-shaped neural network. Frontiers in neuroinformatics **15** (2021) 641600
14. Huang, P., Liu, L., Zhang, Y., Zhong, S., Liu, P., Hong, H., Wang, S., Xie, L., Lin, M., Jiaerken, Y., et al.: Development and validation of a perivascular space segmentation method in multi-center datasets. NeuroImage **298** (2024) 120803
15. Lian, C., Zhang, J., Liu, M., Zong, X., Hung, S.C., Lin, W., Shen, D.: Multi-channel multi-scale fully convolutional network for 3d perivascular spaces segmentation in 7t mr images. Medical image analysis **46** (2018) 106–117
16. Zhuge, M., Gao, D., Fan, D.P., Jin, L., Chen, B., Zhou, H., Qiu, M., Shao, L.: Kaleido-bert: Vision-language pre-training on fashion domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021) 12647–12657
17. Wei, X., Zhang, T., Li, Y., Zhang, Y., Wu, F.: Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 10941–10950
18. Wang, T., Wu, Z., Yao, F., Wang, D.: Graph-based environment representation for vision-and-language navigation in continuous environments. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2024) 8331–8335
19. Liu, S., Hui, T., Huang, S., Wei, Y., Li, B., Li, G.: Cross-modal progressive comprehension for referring segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9) (2021) 4761–4775
20. Tunga, A., Nuthalapati, S.V., Wachs, J.: Pose-based sign language recognition using gcn and bert. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. (2021) 31–40
21. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2) (2021) 203–211
22. Mou, L., Zhao, Y., Fu, H., Liux, Y., Cheng, J., Zheng, Y., Su, P., Yang, J., Chen, L., Frangi, A.F., et al.: Cs2-net: Deep learning segmentation of curvilinear structures in medical imaging. Medical Image Analysis (2020) 101874
23. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. (2022) 574–584

24. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop, Springer (2021) 272–284
25. Sinclair, B., Vivash, L., Moses, J., Lynch, M., Pham, W., Dorfman, K., Marotta, C., Koh, S., Bunyamin, J., Rowsthorn, E., et al.: Perivascular space identification nnunet for generalised usage (pingu). arXiv preprint arXiv:2405.08337 (2024)