CLIP-DSA: Textual Knowledge-Guided Cerebrovascular Diseases Recognition in Multi-View Digital Subtraction Angiography

Qihang Xie[1], Dan Zhang[2,✉], Mengting Liu[3], Jianwei Zhang[4], Ruisheng Su[5], Caifeng Shan[6], and Jiong Zhang[1,✉]

[1] Laboratory of Advanced Theranostic Materials and Technology, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, Ningbo, China
danzhang@nbut.edu.cn;jiong.zhang@ieee.org
[2] School of Cyber Science and Engineering, Ningbo University of Technology, Ningbo, China
[3] Department of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China
[4] USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
[5] Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, Netherlands.
[6] School of Intelligence Science and Technology, Nanjing University, Nanjing, China

Abstract. Digital Subtraction Angiography (DSA) sequences are the gold standard for diagnosing most Cerebrovascular diseases (CVDs). Rapid and accurate recognition of CVDs in DSA sequences helps clinicians make the right decisions, which is important in clinical practice. However, the pathological characteristics of CVDs are numerous and complex, and the spatiotemporal complexity of DSA sequences is high, making the diagnosis of CVDs challenging. Therefore, in this paper, we propose a novel CVDs classification framework CLIP-DSA based on CLIP, a pre-trained vision language model. We aim to utilize textual knowledge to guide the robust classification of common CVDs in multi-view DSA sequences. Specifically, our CLIP-DSA comprises a dual-branch vision encoder and a text encoder. The vision encoder is used to extract features from multi-view sequences, while the text encoder is used to obtain textual knowledge. To optimally harness the temporal information in DSA sequences, we introduce a temporal pooling module that dynamically compresses image features in the time dimension. Additionally, we design a multi-view contrastive loss to enhance the network's image-text representation ability by constraining the image features between two views. In a large dataset with 2,026 patients, the proposed CLIP-DSA achieved an AUC of 90.8% in the CVDs classification. The code is available at this website [1].

Keywords: Vision Language Model, Digital Subtraction Angiography, Cerebrovascular Diseases, Image-Text

[1] https://github.com/jiongzhang-john/CLIP-DSA

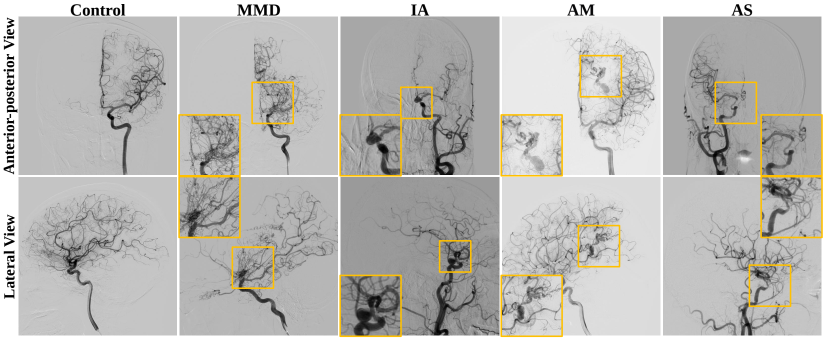

Fig. 1. Schematic diagram of several CVDs, including paired anterior-posterior (AP) and lateral (LA) views. Control: normal/healthy, MMD: Moyamoya disease, IA: intracranial aneurysm, AM: arteriovenous malformation, AS: arterial stenosis.

1 Introduction

Cerebrovascular diseases (CVDs) are among the leading causes of mortality and disability worldwide, imposing significant physical and psychological suffering on patients [1]. These conditions mainly result from abnormalities in cerebrovascular structures. For instance, Moyamoya disease (MMD) is characterized by the appearance of smoke-like vessels at the end of the internal carotid artery, whereas abnormal cerebral artery dilation can lead to aneurysms, as illustrated in Fig. 1. DSA captures a sequence of 2D images over time, depicting the entire angiographic process from arterial to venous phases in both anterior-posterior (AP) and lateral (LA) views. Due to its high spatial and temporal resolution, DSA is considered the gold standard for diagnosing most CVDs [2]. However, the interpretation of DSA heavily relies on visual examination by radiologists, and each patient generates a substantial volume of data, making the process time-consuming and labor-intensive. This complexity increases the risk of missed or incorrect diagnoses, especially for radiologists with limited experience. Therefore, the prompt diagnosis of CVDs is key to faster and more effective treatment to reduce morbidity and mortality [3]. Developing automated methods for this task can enhance diagnostic efficiency and support clinicians in decision-making and treatment planning.

The rapid developments in deep learning have inspired several studies focused on the segmentation and classification of CVDs in DSA sequences. Xie *et al.* [4] and Su *et al.* [5] extracted cerebrovascular structures from DSA sequences using spatiotemporal information. Lei *et al.* [6] proposed a multi-view convolutional neural network (CNN) based on ResNet [7] by combining the AP and LA views to identify the MMD and its hemorrhagic risk. Xu *et al.* [8] introduced a pseudo-3D ResNet to process spatial and temporal information to assess the condition of MMD, including mild, moderate, and severe stages. Similarly, Hu *et al.* [9] utilized 2D CNN, 3D CNN, and BiConvGRU to learn spatiotemporal features of DSA for MMD detection. Mittmann *et al.* [10] presented a new deep learning-based approach to classify thrombus of DSA sequences of patients with acute ischemic stroke. Additionally, some researchers have used deep learning to detect aneurysms [11] and intracranial vessel perforations [12] in DSA sequences. However, these methods can only identify a single disease, while common CVDs have diverse and complex pathological characteristics, leading to insufficient generalization ability for the model. Additionally, the high spatiotemporal complexity of DSA sequences contains a large amount of redundant information, which likely reduces the accuracy of disease recognition.

To address these challenges, we propose to leverage textual knowledge to enhance CVDs recognition. The Contrastive Language-Image Pre-Training (CLIP) model [13] has recently gained attention for its generalization in feature extraction across diverse data types. CLIP-based models have also shown promise in addressing clinical challenges, such as leveraging paired image-text reports to enhance domain-specific knowledge [14]. In this work, we propose CLIP-DSA, a novel framework for recognizing CVDs in multi-view DSA sequences by leveraging CLIP's robust feature learning capabilities. The main contributions are:

(a) We propose a novel framework CLIP-DSA that leverages textual knowledge to guide the robust classification of common CVDs in multi-view DSA sequences. To our knowledge, CLIP-DSA is the first network to utilize textual information from CLIP for DSA-based CVDs recognition.
(b) We introduce a Temporal Pooling Module (TPM), which dynamically compresses image features in time dimension to better integrate the temporal correlations among sequential frames in DSA sequences.
(c) We refine pre-training with contrastive learning at both single- and multi-view levels and introduce a multi-view contrastive loss to improve image-text representation by aligning features across views.

2 Method

2.1 Overall architecture

The proposed CLIP-DSA is trained under the CLIP paradigm, using our constructed multi-view image-text triplets. As shown in Fig. 2, it features a *dual-branch CLIP vision encoder* to extract image features from multi-view DSA sequences and a *frozen CLIP text encoder* to capture textual knowledge. Following the vision encoder, we introduce a temporal pooling module to compress single-view features along the time dimension dynamically. These features are then integrated to form a multi-view artery-level representation, aligned with clinical practice. Each of the three feature types (two single-view and one multi-view) is transformed via linear layers to match the dimensionality of textual knowledge. Finally, similarity scores are computed between each feature and textual knowledge, with a multi-view contrastive loss further constraining the single-view features.

Given an input pair of multi-view DSA sequences S^{ap} and $S^{la} \in \mathbb{R}^{b\times c\times t\times h\times w}$, and their corresponding text embedding $T \in \mathbb{R}^{b\times l}$, a multi-view images-text

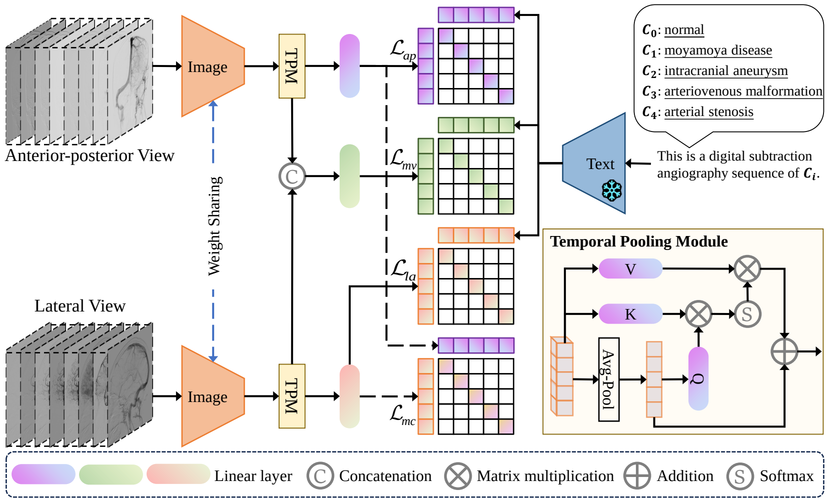

Fig. 2. Overview of the proposed method. $\mathcal{L}_{mc}$ denotes the multi-view contrastive loss.

triplet $\mathcal{D} = \{S^{ap}, S^{la}, T\}$ is constructed. Where b, c, t, h, and w represent the batch sizes, the number of channels, the number of frames, the height, and the width of S, and l is the length of text, respectively. The vision encoder takes S^{ap} and S^{la} as input, while the text encoder is fed with T. **Vision encoder:** The process of sequence encoding has been optimized for training convenience by merging b and t into N, where $N = b \times t$. S^{ap} and S^{la} were encoded using a ResNet50-based encoder $\Phi_v(\cdot)$ respectively:

$$\mathcal{S}^{ap} = \Phi_v(S^{ap}) \in \mathbb{R}^{N\times d}, \mathcal{S}^{la} = \Phi_v(S^{la}) \in \mathbb{R}^{N\times d}, \tag{1}$$

where $\mathcal{S}^{ap}$ and $\mathcal{S}^{la} \in \mathbb{R}^{N\times d}$ denote the features of the two single-view sequences, and d represents the embedding dimension. **Text encoder:** The CLIP text encoder $\Phi_t(\cdot)$ with frozen parameters is implemented to encode the given text embedding T to get the textual knowledge. The complete encoding process can be illustrated as follows:

$$\mathcal{T} = \Phi_t(T) \in \mathbb{R}^{b\times d}, \tag{2}$$

where $\mathcal{T} \in \mathbb{R}^{b\times d}$ represents the textual knowledge.

2.2 Temporal pooling module

It is important to utilize the information on dynamic contrast flow in sequential frames to assist CVDs recognition in clinical practice. Furthermore, the image feature $\mathcal{S}^{ap}$ and $\mathcal{S}^{la} \in \mathbb{R}^{N\times d}$ obtained from the vision encoder cannot be directly calculated similarity with the text feature $\mathcal{T} \in \mathbb{R}^{b\times d}$ extracted from text encoder, necessitating compress in time dimension. Therefore, we designed the Temporal Pooling Module (TPM) to better integrate the temporal correlations among sequential frames in DSA sequences. The TPM can dynamically adjust the weight of feature $\mathcal{S}^{ap}$ and $\mathcal{S}^{la}$ in the time dimension. The structure of TPM is shown in the yellow box of Fig. 2. Specifically, $\mathcal{S}^{ap} \in \mathbb{R}^{N \times d}$ is firstly transposed dimension to $\mathbb{R}^{b \times t \times d}$. Subsequently, $\mathcal{S}^{ap}$ undergoes average pooling to obtain $\mathcal{S}^{ap}_{avg} \in \mathbb{R}^{b \times d}$, which is then subjected to a linear layer to produce f^{ap}_q. Meanwhile, $\mathcal{S}^{ap}$ is passed through two different linear layers to produce f^{ap}_v and f^{ap}_k, respectively. After that, a dynamic weighted feature is obtained through an attention-based mechanism using f^{ap}_q, f^{ap}_k, f^{ap}_v, which is finally added with the average-pooled feature $\mathcal{S}^{ap}_{avg}$ to produce the final enhanced feature $\mathcal{S}^{ap}_{en} \in \mathbb{R}^{b \times d}$. The complete computational process can be illustrated as follows:

$$\mathcal{S}^{ap}_{avg} = AvgPool(\mathcal{S}^{ap}), \tag{3}$$

$$f^{ap}_q = \mathcal{W}(\mathcal{S}^{ap}_{avg}), f^{ap}_k = \mathcal{W}(\mathcal{S}^{ap}), f^{ap}_v = \mathcal{W}(\mathcal{S}^{ap}), \tag{4}$$

$$\mathcal{S}^{ap}_{en} = Softmax(\frac{f^{ap}_q \times f^{ap\intercal}_k}{\sqrt{d}})f^{ap}_v + \mathcal{S}^{ap}_{avg}, \tag{5}$$

where $\mathcal{W}$ represents linear layer. Similarly, $\mathcal{S}^{la}$ is enhanced through the TPM to get the feature $\mathcal{S}^{la}_{en} \in \mathbb{R}^{b \times d}$.

2.3 Multi-view contrastive loss

After processing through the TPM, the single-view features $\mathcal{S}^{ap}_{en}$ and $\mathcal{S}^{la}_{en}$ are concatenated to form the multi-view feature $\mathcal{S}^{mv}_{en}$. These three features are then linearly mapped to obtain features $\mathcal{V}^{ap}$, $\mathcal{V}^{la}$, and $\mathcal{V}^{mv}$, which are in the same dimension as the textual knowledge $\mathcal{T}$. Therefore, the input multi-view image-text triplet $\mathcal{D}$ yields features set $\mathcal{F} = \{\mathcal{V}^{ap}, \mathcal{V}^{la}, \mathcal{V}^{mv}, \mathcal{T}\}$, which is then divided into three subsets: $\mathcal{F}^{ap} = \{\mathcal{V}^{ap}, \mathcal{T}\}$, $\mathcal{F}^{la} = \{\mathcal{V}^{la}, \mathcal{T}\}$, $\mathcal{F}^{mv} = \{\mathcal{V}^{mv}, \mathcal{T}\} \in \mathbb{R}^{b \times d}$, corresponding to AP view, LA view, and multi-view level, respectively. In each subset, the paired image and text features are considered positive samples for each other, while the rest are negative samples. The cosine similarity matrix is calculated for each subset. For the AP view subset, we measure the similarity inter-samples using Eq. (6) following CLIP [13], termed as SIM^{ap}_{v2t} and $SIM^{ap}_{t2v} \in \mathbb{R}^{b \times b}$, where ϵ is a learnable parameter. Note that the learnable parameter ϵ differs across the various loss functions.

$$SIM^{ap}_{v2t} = \frac{\mathcal{V}^{ap} \times \mathcal{T}^{\intercal}}{\epsilon}, SIM^{ap}_{t2v} = SIM^{ap\,\intercal}_{v2t}, \tag{6}$$

then we calculate the contrastive loss $\mathcal{L}_{ap}$ of the AP view subset, and $\mathcal{L}_{ap}$ can be defined as:

$$\mathcal{L}_{ap} = \frac{CE(SIM^{ap}_{v2t}, Y) + CE(SIM^{ap}_{v2t}, Y)}{2}, \tag{7}$$

where CE and $Y \in \mathbb{R}^{b \times b}$ represent the cross entropy loss and the one-hot labels. Similarly, we can calculate $\mathcal{L}_{la}$ and $\mathcal{L}_{mv}$ for LA view and multi-view according to Eq. (6) and (7).

Additionally, we expect that the AP and LA views, as different perspectives of the same angiography, should exhibit a high degree of similarity. To this end, we utilize a multi-view contrastive loss to constrain the image features from these two views, thereby enhancing the image-text representation capability of CLIP-DSA. The multi-view contrastive loss $\mathcal{L}_{mc}$ can be also calculated by Eq. (6) and (7). Therefore, the overall loss is formulated as follows:

$$\mathcal{L}_{total} = \frac{\mathcal{L}_{ap} + \mathcal{L}_{la} + \mathcal{L}_{mv} + \mathcal{L}_{mc}}{4} \tag{8}$$

3 Experiments

3.1 Dataset and implementation details

The dataset used in this study was collected from *** hospital, comprising 2,897 arteries with paired DSA sequences from 2,026 patients, totaling 5,794 sequences. Based on clinical reports, arteries were labeled as Control (846), MMD (773), IA (320), AM (411), and AS (547). The dataset was split into training (2,320 arteries: Control 678, MMD 619, IA 256, AM 329, AS 438) and test (577 arteries) sets at an approximate 4:1 ratio, with the training set further divided into training and validation subsets using the same ratio. All DSA sequences were obtained from ethically approved studies with written informed consent, following the Declaration of Helsinki. The text data consists of sentences formed by disease labels, as shown in Fig. 2. Our method was implemented in PyTorch using NVIDIA GeForce RTX 4090 GPUs. Model training employed AdamW [15] with a weight decay of 1e-5 for 200 epochs. A polynomial decay strategy adjusted the learning rate from 1e-4 to 1e-6. The batch size was set to 16, images were resized to 224×224 pixels, and 16 frames were resampled per sequence. Common data augmentation techniques, including horizontal and vertical flipping, were applied. CLIP [13] and our CLIP-DSA used a pre-trained ResNet50 [7] as the vision encoder, while the text encoder was a pre-trained Transformer.

3.2 Comparison with state-of-the-art methods

To assess the performance of the proposed CLIP-DSA, we compared it with several various state-of-the-art methods, including 2D methods (ResNet50 [7], Swin_s [16]), Recurrent Neural Network (RNN) (Mittmann *et al.* [10], Hu *et al.* [9]), 3D methods (R3D [17], S3D [18], MViT_s [19], Swin3D_s [20]) and vision-language models (VLM) (HowTo100M [21], CLIP [13]). Pre-trained weights were loaded if available for the compared methods. Additionally, to ensure the fairness of the comparative experiments, these methods also incorporated a dual-branch image encoder with shared weights, along with three classification heads corresponding to the AP view, LA view, and multi-view, similar to CLIP-DSA. We used the accuracy (ACC), F1 score (F1), and area under the ROC curve (AUC) to evaluate the classification performance. Note that in Table 1 and Table 2, the

Table 1. Performance comparisons for cerebrovascular disease recognition.

	Methods	ACC(%)			F1(%)			AUC(%)		
		AP	LA	MV	AP	LA	MV	AP	LA	MV
2D	ResNet50 [7]	78.0	79.7	82.3	75.4	78.2	80.7	85.2	87.3	88.0
	Swin_s [16]	75.7	77.3	82.0	73.3	76.0	79.6	83.7	85.8	87.2
RNN	Mittmann *et al.* [10]	77.4	79.7	82.0	75.2	77.7	79.4	85.8	86.4	87.6
	Hu *et al.* [9]	54.8	50.8	56.0	49.1	47.6	53.2	69.7	67.6	72.2
3D	R3D [17]	77.3	78.3	80.4	75.8	76.7	78.7	85.7	85.5	86.7
	S3D [18]	74.0	77.1	79.4	71.4	75.0	77.9	82.9	85.5	87.3
	MViT_s [19]	73.0	73.0	77.3	71.0	71.1	76.0	83.3	83.4	85.8
	Swin3D_s [20]	75.0	77.3	80.4	72.4	74.1	78.5	83.5	84.5	87.1
VLM	HowTo100M [21]	66.9	70.9	72.4	64.9	70.4	71.9	80.5	83.0	85.1
	CLIP [13]	75.9	78.7	81.3	74.0	76.5	79.3	84.0	85.3	87.1
	CLIP-DSA	-	-	**84.1**	-	-	**83.3**	-	-	**90.8**

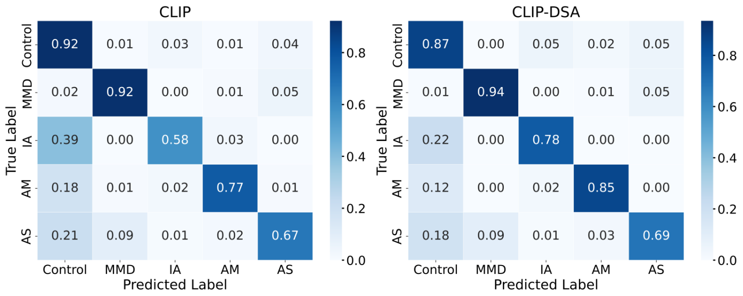

Fig. 3. The confusion matrices of CLIP and CLIP-DSA for recognizing cerebrovascular diseases from multi-view DSA sequences.

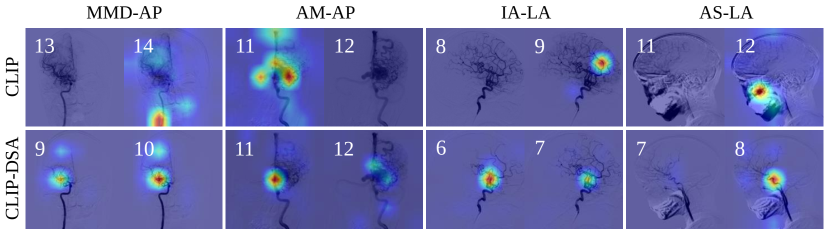

Fig. 4. The GradCAM diagrams of CLIP and CLIP-DSA, including MMD, AM, IA, and AS. The number represents the frame index, as the most decisive frame varies across different methods.

AP and LA columns indicate that the network uses only single-view sequences as input, while MV represents the use of multi-view sequences.

The comparison results are presented in Table 1. From the quantitative analysis, it is evident that the proposed CLIP-DSA outperforms other methods. Since

Table 2. Ablation results of our CLIP-DSA for cerebrovascular disease recognition.

Components		ACC(%)			F1(%)			AUC(%)		
Pooling Module	$\mathcal{L}_{mc}$	AP	LA	MV	AP	LA	MV	AP	LA	MV
avg_pool		75.9	78.7	81.3	74.0	76.5	79.3	84.0	85.3	87.1
✓		77.1	78.7	82.0	75.2	77.0	80.0	85.1	86.3	87.8
avg-pool	✓	-	-	83.5	-	-	82.3	-	-	89.4
max-pool	✓	-	-	81.7	-	-	80.5	-	-	88.8
lstm	✓	-	-	81.3	-	-	80.4	-	-	89.1
gru	✓	-	-	79.4	-	-	78.0	-	-	86.9
✓	✓	-	-	**84.1**	-	-	**83.3**	-	-	**90.8**

the multi-view contrastive loss requires constraining features from both the AP and LA views, metrics cannot be calculated for a single view independently. However, our method achieves the highest performance in terms of ACC, F1 score, and AUC at the multi-view level. Specifically, CLIP-DSA improves upon ResNet50 by 1.8%, 2.6%, and 2.8% in ACC, F1, and AUC, respectively; surpasses Swin3D_s by 3.7%, 4.8%, and 3.7%; and exceeds the CLIP backbone by 2.8%, 4.0%, and 3.7%, demonstrating its effectiveness and superiority with statistical significance ($p<0.05$). Table 1 shows that multi-view recognition for cerebrovascular disease outperforms single-view approaches. Interestingly, 2D methods outperform RNN and 3D methods, likely because lesions are typically visible only in a few middle frames, and improper handling of temporal data can negatively impact performance. Additionally, a normalized confusion matrix of CLIP and our CLIP-DSA on the test set is shown in Fig. 3. This indicates that our CLIP-DSA achieves higher accuracy in recognizing MMD, IA, AM, and AS. Compared to CLIP, our method demonstrates a stronger capability for image-text representation. To further compare the two models, we used Grad-CAM[22] to generate class activation maps for CLIP-DSA and CLIP. It can be seen that CLIP-DSA focuses more on the lesion area compared to CLIP, as shown in Fig. 4.

3.3 Ablation study

To investigate the effectiveness of each component in the proposed CLIP-DSA, we conducted the ablation studies in Table 2. We employed the CLIP as the backbone, systematically reintegrating each component to conduct comprehensive ablation studies. The $3rd$ row of Table 2 is the backbone CLIP, while $4th$ and $5th$ row represent adding TPM and $\mathcal{L}_{mc}$, respectively. They all contribute to the improvement of the CVDs recognition effect. Notably, adding the $\mathcal{L}_{mc}$ has the most significant improvement, i.e., 2.2% in ACC, 3.0% in F1 score, and 2.3% in AUC. This suggests that the features of the AP and LA views hold highly similar information and constraining image features of the two views can further enhance the network's image-text representation capability. Additionally, to further validate the effectiveness of TPM, we replaced it with max pooling, LSTM, and GRU, as shown in rows 6 to 8 of Table 2. Overall, CLIP-DSA achieves superior performance by combining the proposed components.

4 Conclusion

In this paper, we have proposed a novel framework, CLIP-DSA, the first network that leverages textual information from CLIP for CVD recognition in multi-view DSA sequences. It incorporates a temporal pooling module and a multi-view contrastive loss to enhance the network's image-text representation, improving CVD recognition accuracy. Extensive experiments validate the effectiveness of our approach, highlighting its potential for clinical application.

Acknowledgments. 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. Mensah, G.A., Fuster, V., Murray, C.J., Roth, G.A., of Cardiovascular Diseases, G.B., Collaborators, R.: Global burden of cardiovascular diseases and risks, 1990-2022. Journal of the American College of Cardiology **82**(25) (2023) 2350–2473
2. Hess, C.P.: Imaging in cerebrovascular disease. In: Molecular, Genetic, and Cellular Advances in Cerebrovascular Diseases. World Scientific (2018) 1–23
3. Hussein, R., Zhao, M.Y., Shin, D., Guo, J., Chen, K.T., Armindo, R.D., Davidzon, G., Moseley, M., Zaharchuk, G.: Multi-task deep learning for cerebrovascular disease classification and mri-to-pet translation. In: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE (2022) 4306–4312
4. Xie, Q., Zhang, D., Mou, L., Wang, S., Zhao, Y., Guo, M., Zhang, J.: Dsnet: A spatio-temporal consistency network for cerebrovascular segmentation in digital subtraction angiography sequences. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2024) 199–208
5. Su, R., van der Sluijs, P.M., Chen, Y., Cornelissen, S., van den Broek, R., van Zwam, W.H., van der Lugt, A., Niessen, W.J., Ruijters, D., van Walsum, T.: Cave: Cerebral artery–vein segmentation in digital subtraction angiography. Computerized Medical Imaging and Graphics **115** (2024) 102392
6. Lei, Y., Zhang, X., Ni, W., Yang, H., Su, J.B., Xu, B., Chen, L., Yu, J.H., Gu, Y.X., Mao, Y.: Recognition of moyamoya disease and its hemorrhagic risk using deep learning algorithms: sourced from retrospective studies. Neural Regeneration Research **16**(5) (2021) 830–835
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
8. Xu, J., Wu, J., Lei, Y., Gu, Y.: Application of pseudo-three-dimensional residual network to classify the stages of moyamoya disease. Brain Sciences **13**(5) (2023) 742
9. Hu, T., Lei, Y., Su, J., Yang, H., Ni, W., Gao, C., Yu, J., Wang, Y., Gu, Y.: Learning spatiotemporal features of dsa using 3d cnn and biconvgru for ischemic moyamoya disease detection. International Journal of Neuroscience **133**(5) (2023) 512–522

10. Mittmann, B.J., Braun, M., Runck, F., Schmitz, B., Tran, T.N., Yamlahi, A., Maier-Hein, L., Franz, A.M.: Deep learning-based classification of dsa image sequences of patients with acute ischemic stroke. International journal of computer assisted radiology and surgery **17**(9) (2022) 1633–1641
11. Duan, H., Huang, Y., Liu, L., Dai, H., Chen, L., Zhou, L.: Automatic detection on intracranial aneurysm from digital subtraction angiography with cascade convolutional neural networks. Biomedical engineering online **18** (2019) 1–18
12. Su, R., van der Sluijs, M., Cornelissen, S.A., Lycklama, G., Hofmeijer, J., Majoie, C.B., van Doormaal, P.J., Van Es, A.C., Ruijters, D., Niessen, W.J., et al.: Spatio-temporal deep learning for automatic detection of intracranial vessel perforation in digital subtraction angiography during endovascular thrombectomy. Medical image analysis **77** (2022) 102377
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR (2021) 8748–8763
14. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training. Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations. (2017)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. (2021) 10012–10022
17. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2018) 6450–6459
18. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). (2018) 305–321
19. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. (2021) 6824–6835
20. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2022) 3202–3211
21. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision. (2019) 2630–2640
22. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. International journal of computer vision **128** (2020) 336–359