Multi-grained contrastive representation learning for label-efficient lesion segmentation and onset time classification of acute ischemic stroke

Jiarui Sun, Yuhao Liu, Yan Xi, Gouenou Coatrieux, Jean-Louis Coatrieux, Xu Ji, Liang Jiang, Yang Chen

Multi-Grained Contrastive Representation Learning for Label-Efficient Lesion Segmentation and Onset Time Classification of Acute Ischemic Stroke

Jiarui Sun[a], Yuhao Liu[b], Yan Xi[c], Gouenou Coatrieux[d], Jean-Louis Coatrieux[e,f], Xu Ji[a,g,*], Liang Jiang[h,*], Yang Chen[a,i]

[a]*Laboratory of Image Science and Technology, Southeast University, Nanjing 210096, China*
[b]*Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong*
[c]*Jiangsu First-Imaging Medical Equipment Co., Ltd., Nanjing 210009, China*
[d]*IMT Atlantique, Inserm, LaTIM UMR1101, Brest 29000, France*
[e]*Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, F-35000 Rennes, France*
[f]*Centre de Recherche en Information Biomédicale Sino-francais, 35042 Rennes, France*
[g]*Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 210096, China*
[h]*Department of Radiology, Nanjing First Hospital, Nanjing Medical University, Nanjing 210006, China*
[i]*Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, Nanjing 210096, China*

Keywords: Acute ischemic stroke analysis, Multi-modal MRI imaging, Multi–grained contrastive learning, Prior representation

A B S T R A C T

Ischemic lesion segmentation and the time since stroke (TSS) onset classification from paired multi-modal MRI imaging of unwitnessed acute ischemic stroke (AIS) patients is crucial, which supports tissue plasminogen activator (tPA) thrombolysis decision-making. Deep learning methods demonstrate superiority in TSS classification. However, they often overfit task-irrelevant features due to insufficient paired labeled data, resulting in poor generalization. We observed that unpaired data are readily available and inherently carry task-relevant cues, but are less often considered and explored. Based on this, in this paper, we propose to fully excavate the potential of unpaired unlabeled data and use them to facilitate the downstream AIS analysis task. We first analyse the utility of features at the varied grain and propose a multi-grained contrastive learning (MGCL) framework to learn task-related prior representations from both coarse-grained and fine-grained levels. The former can learn global prior representations to enhance the location ability for the ischemic lesions and perceive the healthy surroundings, while the latter can learn local prior representations to enhance the perception ability for semantic relation between the ischemic lesion and other health regions. To better transfer and utilize the learned task-related representation, we designed a novel multi-task framework to simultaneously achieve ischemic lesion segmentation and TSS classification with limited labeled data. In addition, a multi-modal region-related feature fusion module is proposed to enable the feature correlation and synergy between multi-modal deep image features for more accurate TSS decision-making. Extensive experiments on the large-scale multi-center MRI dataset demonstrate the superiority of the proposed framework. Therefore, it is promising that it helps better stroke evaluation and treatment decision-making.

*Corresponding author

e-mail: xuji@seu.edu.cn (Xu Ji), jiangliang0402@163.com (Liang Jiang)

1. Introduction

Stroke is a common cerebrovascular disease with the fifth leading cause of death (Vijayan and Reddy, 2016). Acute ischemic stroke (AIS) is the most common subtype, which leads to 2.7 million deaths worldwide every year (Benjamin et al., 2019). Treatment of AIS is strictly dependent on the time since stroke onset (TSS). According to the AIS treatment guidelines, TSS within 4.5 hours is the golden time window of tissue plasminogen activator (tPA) thrombolysis due to increased hemorrhage risk when administered beyond that time interval (Campbell et al., 2019). However, approximately 30% of AIS patients are excluded from tPA treatment because of unknown TSS while they may be within the time window of the tPA thrombolysis (Moradiya and Janjua, 2013). Thus, the guidelines from American Stroke Association (ASA) recommend using paired multi-modal MRI imaging to classify TSS to determine thrombolysis eligibility of the unwitnessed AIS patients (Powers et al., 2019).

The mismatch pattern of the diffusion-weighted imaging (DWI)-fluid attenuated inversion recovery (FLAIR) imaging is the most common way of classifying TSS (Powers et al., 2019). The DWI-FLIAR mismatch pattern is based on the fact that ischemic lesions are immediately visible on the DWI imaging, but it usually takes about 4 hours to find the ischemic lesion on FLAIR imaging (Thomalla et al., 2011; Ebinger et al., 2010; Emeriau et al., 2013). As depicted in Fig.1, it can be observed that the high-intensity signal on DWI imaging is not visible in the corresponding location of FLAIR imaging, which means DWI-positive FLAIR-negative lesions. Therefore, TSS can be classified via the DWI-FLAIR mismatch pattern. Besides, current studies have demonstrated MR perfusion-weighted imaging (PWI) contains information encoding TSS (Murphy et al., 2007; McLeod et al., 2011; Thomalla and Gerloff, 2015; Jiang et al., 2024). Moreover, related clinical studies have shown that about 80% of AIS patients have ischemic penumbra caused when TSS is less than 3 hours (Davis et al., 2008). As described in Fig.1, it can be observed that the high-intensity signal regions on DWI and PWI imaging are not matched in shape and size, which means the DWI-PWI mismatch pattern. Therefore, TSS can also be more strictly classified via the DWI-PWI mismatch pattern. Especially, under the guidance of the PWI-DWI mismatch pattern, tPA thrombolysis treatment may be more reliable and have a better prognosis (Wolman et al., 2018). While the two mismatch patterns are the current advanced method for clinically determining TSS for unwitnessed AIS patients, computing mismatch using paired DWI-FLAIR or DWI-PWI imaging is a difficult and time-consuming task that requires extensive clinical training. Thus, assessing this mismatch is naturally subject to high variability across multiple inter-observers and radiologists (Ziegler et al., 2012; Thomalla et al., 2011; Galinovic et al., 2014). Besides, it may miss some individuals who would benefit from tPA thrombolysis treatment because of overly strict mismatch conditions (Odland et al., 2015).

Data-driven methods (Ho et al., 2019; Thomalla et al., 2009; Zhu et al., 2021; Zhang et al., 2021; Jiang et al., 2022) demonstrate great potential for AIS analysis due to the high capability in excavating representative features. These methods utilize

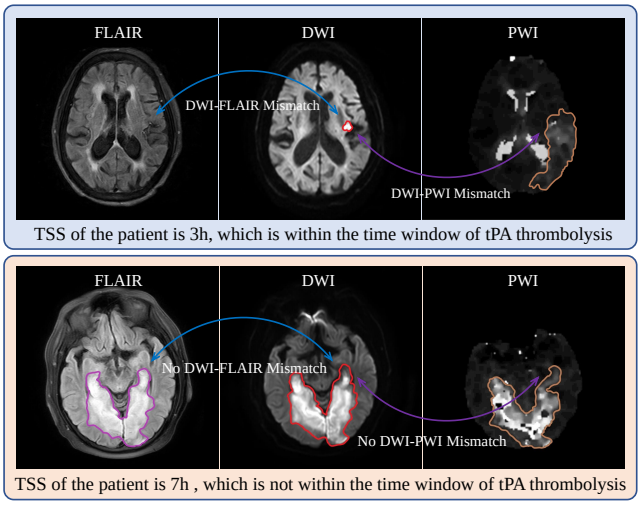

Fig. 1: Paired multi-modal MRI sequences of two AIS patients are presented. The ischemic lesions are delineated in different colors including purple for FLAIR, red for DWI, and orange for PWI. In the above figure, there is the presence of DWI-FLAIR and DWI-PWI mismatches (ie., TSS of the given AIS patient is 3h), and the absence of DWI-FLAIR and DWI-PWI mismatches (ie., TSS of the given AIS patient is 7h).

hand-crafted, radionics, or deep learning-driven features extracted from multi-modal MRI images, and these features are incorporated into machine learning models for TSS classification. As for the data-driven methods, the location information of ischemic lesions is crucial because feature extraction typically relies on lesion regions and healthy surroundings (Bang, 2011). On the other hand, current studies (Murphy et al., 2007; McLeod et al., 2011; Thomalla and Gerloff, 2015; Ho et al., 2019) demonstrate that PWI images contain important information encoding TSS, and combining it with FLAIR and DWI imaging can in turn boost and improve TSS classification performance. Inspired by the above, our work focuses on employing paired multi-modal MRI imaging (DWI, FLAIR and PWI sequences) for the comprehensive AIS analysis, including the ischemic lesion segmentation and TSS classification of the unwitnessed AIS patients. However, two inherent limitations hinder the development and cause performance bottlenecks. **Limitation 1**: Limited by the urgency of AIS onset and medical conditions, collecting large-scale paired multi-modal MRI imaging, including DWI, FLAIR and PWI sequences, is very difficult. Therefore, existing data-driven deep learning methods based on image- (Zabihollahy et al., 2020; Pedersen et al., 2020) or lesion-level (Yu et al., 2016) tend to overfit task-irrelevant features due to the interference of the chaotic background or inefficient edge feature extraction for the lesion regions, which leads to poor generalization ability. (Xu et al., 2020; Chen et al., 2021; Kong et al., 2022). **Limitation 2**: Current TSS classification methods usually rely on the simple fusion of the extracted multi-modal image features and typically fail to explicitly consider the clinal diagnostic knowledge in the deep features fusion process (*e.g.,* DWI-FLAIR or DWI-PWI mismatch patterns in Fig.1), resulting in the lower feature utilization.

For limitation 1. For image analysis tasks, emerging contrastive representation learning shows great potential in exploit-

ing massive unlabeled data, which helps models obtain a better generalization ability with limited annotation data (Azizi et al., 2021; Yang et al., 2022; Li et al., 2020; Han et al., 2022; Wu et al., 2022). We reasonably made use of the fact: compared to paired multi-modal MRI data, a large amount of the unpaired data is readily available yet underutilized and under-explored. Thus, we proposed a novel multi-grained contrastive learning (MGCL) framework to learn task-related prior representations via developing large-scale unpaired data. As a result, learned representations can boost efficient utilization of the limited segmentation and classification labels of paired multi-modal MRI data. Especially, MGCL learns and transfers prior representations via two cascade stages. In stage 1, models learn task-related prior representations from the massive unlabeled unpaired data based on two task-specific contrastive learning versions: a) Coarse-grained version encourages the models to learn global prior representations to enhance the location ability for the ischemic lesions while perceiving the healthy surroundings, which helps capture the task-related features in AIS analysis. b) Fine-grained version encourages the models to learn local prior representations to enhance the perception ability for semantic relation between the ischemic lesion and other health regions, which helps supplement the lesion details. In stage 2, the learned task-specific prior representations are reasonably transferred into a designed multi-task learning architecture, which comprehensively improves the performance of the ischemic lesion segmentation and TSS classification with limited paired MRI data.

For limitation 2. We propose a multi-modal region-related feature fusion (MRFF) module to adequately consider the feature relationship between paired multi-modal MRI images with sub-region as the basic unit, which can explicitly integrate the diagnostic knowledge of two mismatch patterns into TSS decision-making. Especially, it can capture the correlation and synergy among corresponding image regions of the paired multi-modal MRI sequences, which improves generalization ability by alleviating overfit task-irrelevant features from feature correlation calculation of non-corresponding regions. Finally, the calculated feature correlation is mapped into a multi-modal fusion feature to support efficient TSS classification.

In general, our contributions include the following:

- For the first time, we propose a multi-grained contrastive learning framework based on two cascade stages. In stage 1, the model learns task-related prior representations to exploit massive unlabeled MRI data. In stage 2, the learned prior representations are reasonably transferred into the designed multi-task learning architecture.
- We propose two task-specific contrastive feature enhancement strategies for the representation learning in MGCL. The coarse-grained version learns global prior representations to enhance the location ability for the ischemic lesions and perceives the healthy surroundings, while the fine-grained version learns local prior representations to enhance the perception ability for semantic relation between the ischemic lesion and other health regions.
- We propose a multi-modal region-related feature fusion module to adequately capture the feature relationship between multi-modal MRI images, which can explicitly integrate the clinical diagnostic knowledge of multiple mismatch patterns into TSS classification.
- We construct a large-scale multi-center multi-modal (DWI, FLAIR and PWI) MRI dataset for the AIS analysis, which includes the labeled part and the unlabeled part. The labeled part contains 327 paired multi-modal MRI images with the patient-level TSS classification label (TSS<4.5hrs) and strict pixel-level ischemic lesion annotations. The unlabeled part contains massive unpaired MRI images.
- Extensive experiments also show the superiority of the proposed framework. The data will be made public at https://github.com/JiaRuiS/MGCL.

2. Related Work

In this section, we review automatic TSS classification and self-supervised contrastive learning (SSCL) literature that are closely relevant to our work.

2.1. Automatic TSS classification

Machine learning methods (Zhu et al., 2021; Jiang et al., 2022; Ho et al., 2019; Lee et al., 2020; Jiang et al., 2024) extract image features by hand-crafted, radiomics or deep learning (DL), and then utilize classifiers (eg., support vector machine, Bayesian classifier, and logistic models) to achieve the TSS classification. Zhu (Zhu et al., 2021) and Jiang (Jiang et al., 2022) et al. use DL methods to calculate the regions of interest (ROI) of ischemic lesions from DWI and FLAIR images, extract radiomics feature, and finally classify TSS by voting from the results of multiple machine learning classifiers. Ho et al. (Ho et al., 2019) proposed a deep autoencoder model to extract hidden representations from the PWI images and combine the baseline features of DWI and FLAIR images to jointly classify TSS. Jiang et al. (Jiang et al., 2024) proposed a segmentation-classification model to automatically identify stroke within 4.5 h based on DWI and PWI fusion images. With the rapid development of DL in medical image analysis, DL-based methods reduce the trouble of designing task-related feature extraction methods and achieve better performance (Zhang et al., 2021; Polson et al., 2022). Relying on the feature extraction ability obtained from the pre-training task of stroke detection, Zhang et al. (Zhang et al., 2021) employed the modified 2D and 3D CNN architectures to classify TSS. Polson et al. (Polson et al., 2022) propose a novel 2.5D CNN based on the improved ResNet-34 (He et al., 2016). It incorporates the inter-slice information into 2D CNNs to extract the different modal features and aggregate each source feature via a multi-modal information fusion manner to improve TSS classification performance.

Unlike the natural image classification tasks, the dataset size of the TSS classification task is usually smaller because of medical conditions. Thus, DL-based methods are easier to overfit, which leads to poor generalization ability. Besides, current TSS classification methods typically fail to explicitly integrate the diagnostic information including DWI-FLAIR or DWI-PWI mismatch patterns, which may impede further performance improvement.

2.2. Self-Supervised Contrastive Learning

Self-supervised learning (SSL) aims to develop massive unlabeled data. The key to SSL is to design reasonable proxy tasks to generate supervisory signals for unlabeled data (Kolesnikov et al., 2019). SSL is usually categorized into generative and discriminative approaches depending on the proxy tasks (Liu et al., 2021). As a representative discriminative approach, emerging contrastive learning (CL) shows great potential. The core idea is to contrast the similarity of sample pairs via a contrastive loss, pull semantically nearby image samples (positive pairs) closer, and push dissimilar image samples (negative pairs) apart. Self-supervised contrastive learning (SSCL) provides a standard paradigm for the image analysis field: the model learns image representations with massive unlabeled data through pre-training. Then, the pre-trained model can be used as the initialization for improving the performance of the downstream supervised task (ie ., image classification (Azizi et al., 2021; Misra and Maaten, 2020) and image segmentation (Wu et al., 2022; Araslanov and Roth, 2021)).

*2.2.1. **SSCL for natural images***

In the natural image analysis field, two well-known methods including MOCO (He et al., 2020) and SimCLR (Chen et al., 2020) were first proposed, which learn knowledge representations from large-scale unlabeled data by contrastive learning to boost the performance of downstream supervised tasks. The results suggested that they significantly narrowed the gap in downstream task performance between self-supervised learning and fully-supervised learning. They believe that the magnitude of negative pairs plays an important role in performance improvement. Soon after, Grill and Chen et al. (Grill et al., 2020; Chen and He, 2021) proposed BYOL and SimSiam, which also demonstrated that negative pairs are not necessary for contrastive learning.

Contrastive learning shows great potential in natural image analysis. However, the above methods are all specially designed for natural images and do not take into account the domain-specific knowledge (eg., anatomical structure knowledge, topology knowledge) of medical images.

*2.2.2. **SSCL for medical images***

Compared to natural image analysis, the annotation of medical images requires a large amount of domain-specific knowledge to guide, which makes it very expensive. However, it is easier to collect a large amount of unlabeled data than manually labeling an accurate large-scale dataset. Therefore, it is quite necessary to develop the SSCL methods for medical image analysis. Zeng et al. (Zeng et al., 2021) proposed a positional contrastive learning (PCL) method, which generated contrastive data pairs by leveraging the position information of different slices in volumetric medical images. They also proved the effectiveness on the downstream segmentation tasks based on several CT or MRI datasets. Considering that slice-level contrastive learning may lack distinctive representations of local regions, Chaitanya and Hu et al. (Chaitanya et al., 2020; Hu et al., 2021) propose to capture the global and local representations, which find domain-specific and problem-specific cues. Besides, Chaitanya et al. (Chaitanya et al., 2023) provides a new perspective on how to learn the semantic-guided local representations by contrastive learning for improving segmentation performance with very limited annotation.

Considering the importance of domain-specific prior knowledge for medical image tasks, current methods usually exists a certain semantic knowledge gap between the upstream supervisory signals generated for unlabeled data and the downstream expert annotation (ie., Learning the position relationship representations between different patches on upstream unlabeled data may not be very helpful for the downstream medical image segmentation task). Thus, the learned representations may fail to fully motivate the downstream task performance with limited labeled data sufficiently.

3. Methodology

Fig.2 depicts the pipeline of the proposed MGCL framework for AIS analysis including ischemic lesion segmentation (Task 1) and TSS classification (Task 2) via two cascade stages. Stage 1 learns task-related prior representations to explore large-scale unpaired data via the multi-grained version CL. Stage 2 reasonably transfers learned prior representations to a multi-task learning architecture to efficiently promote the fine-tuning process on Task 1 and Task 2 with limited paired data.

3.1. Overall Architecture

Before stage 1, three independent u-net (Ronneberger et al., 2015) networks (for DWI, FLAIR and PWI sequences) were firstly trained respectively to construct the supervision signal for massive unpaired unlabeled data. Especially, these trained models were employed to generate two types of pseudo-labels for describing ischemic lesions on the slice-level and patch-level. In stage 1, the models gain prior representations based on two task-specific CL versions. The coarse-grained CL version employs the slice-level supervisory signal to incentivize the models to learn the global prior representation. The fine-grained CL version employs the patch-level supervisory signal to incentivize the models to learn local prior representation. In stage 2, the implementation of Task 1 and Task 2 follows the multi-task learning architecture. By transferring learned task-specific prior representations to this multi-task learning architecture, which efficiently promotes the fine-tuning process on Task 1 and 2 with limited paired data. Especially, MRFF computes the correlation and synergy between the multi-modal deep image features on the feature level and makes the feature correlation between different MRI sequences mapped into a fused multi-modal feature to support final TSS decision-making.

3.2. Multi-Grained Contrastive Learning

*3.2.1. **Supervisory Signal Construction***

First, three independent u-net networks (Ronneberger et al., 2015) including an encoder and a decoder are trained separately for the three MRI sequences via a supervised learning fashion respectively (refer to Section 4.2). Then, utilizing the trained u-net networks, the segmentation pseudo labels can be generated by inferencing large-scale unpaired unlabeled data of the three

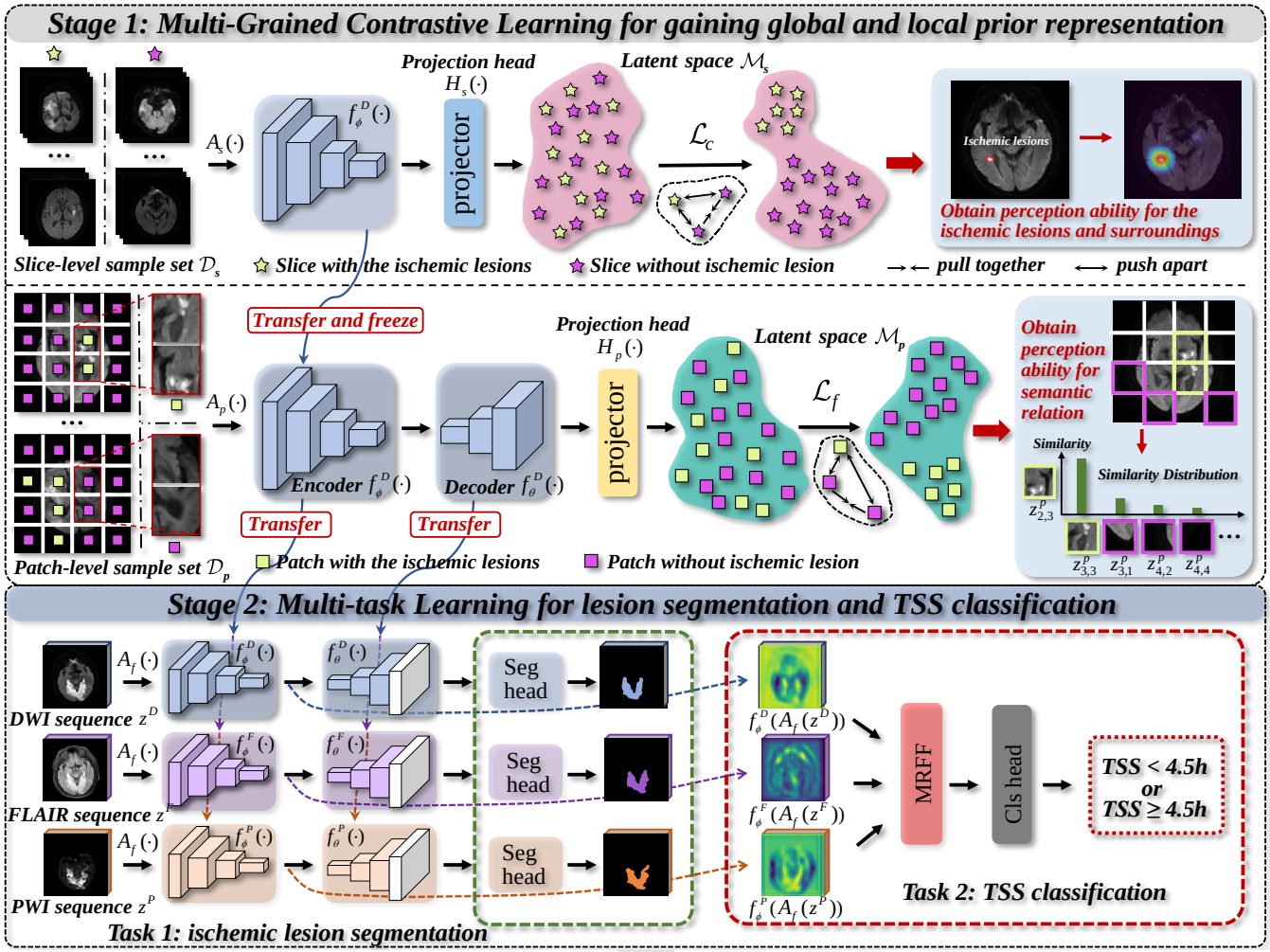

Fig. 2: The schematic illustration of the proposed Multi-Grained Contrastive Learning (MGCL) framework. It includes two cascade stages: (a) Stage 1 sequentially performs coarse and fine-grained contrastive learning to gain global and local prior representation. In this stage, each in the three different MRI modals (DWI, FLAIR and PWI) maintains an independent set including an encoder and a decoder separately. (b) Stage 2 utilizes learned representations from stage 1 as a pertinent initialization for the given multi-task learning architecture to transfer task-specific prior knowledge into the ischemic lesion segmentation (Task 1) and TSS classification (Task 2) tasks. $A_s(\cdot)$, $A_p(\cdot)$, $A_f(\cdot)$ denote the online data augmentation of coarse-grained version, fine-grained version, and multi-task learning, respectively. $z^p_{u,v}$ means the patch-level embedding at the u-th column and v-th row of the given slice.

MRI sequences respectively. Pseudo labels can indicate the lesion region on each slice of a given MRI sequence. Thus, these pseudo labels can determine whether the ischemic lesion occurs at three different sample levels including slice-level, patch-level, and pixel-level. Naturally, it also can be used as the supervisory signal to indicate whether the ischemic lesion occurs on three sample levels. Especially, the pixel-level supervision signals are excluded, because they cause more noise than the slice-level and patch-level. Then, the slice-level and patch-level supervisory signals are reserved to guide the representations learning of different versions. Finally, on the unlabeled part of every MRI sequence, two supervisory signal sets on two different levels are constructed: (a) $\mathcal{D}_s := \left\{(x_i^s, y_i^s)\right\}_{i=1}^{M}$ consists of the M slice-level sample-label pairs. (b) $\mathcal{D}_p := \left\{(x_{u,v}^p, y_{u,v}^p)\right\}_{u,v=1}^{N}$ consists of the N patch-level sample-label pairs. For each of the three MRI sequences, an independent set including an encoder and a decoder is always maintained in stage 1, which can better pay attention to and retain the modal specificity between different sequences during the representation learning of the MGCL.

3.2.2. *Coarse-Grained Version*

As described in stage 1 of Fig.2, utilizing slice-level supervisory signal set $\mathcal{D}_s$, the coarse-grained CL version aims to make the models gain the global prior representations by a learnable encoder way. Firstly, the slice x_i^s in $\mathcal{D}_s$ is projected to latent space $\mathcal{M}_s$ as a $L2$-normalized d-dimension embedding v_i^s, which can be defined as:

$$v_i^s = H_s(f_\phi(A_s(x_i^s))) \,, \tag{1}$$

where $A_s(\cdot)$ is the online data augmentation of coarse-grained version, $f_\phi(\cdot)$ is the learnable encoder, and $H_s(\cdot)$ is the projection head in the coarse-grained version. To measure the distance between the given two slices in $\mathcal{M}_s$, cosine similarity is utilized to calculate the similarity between the given two slice embeddings

v_i^s and v_j^s:

$$s(v_i^s, v_j^s) = \frac{(v_i^s)^{\mathrm{T}} v_j^s}{\|v_i^s\| \|v_j^s\|}, \tag{2}$$

where $\|v_i^s\|$ represents the $L2$ norm of the given embedding v_i^s.

Given a randomly sampled batch, $\mathcal{I}$ denotes the slice indexes. Set x_i^s as the anchor sample, $\mathcal{P}(i) := \{j \in \mathcal{I} \mid y_j^s = y_i^s,\ j \neq i\}$ represents the set of indexes for all augmented positive samples that are with the same label as y_i^s. Naturally, the negative samples are these augmented slices with different labels to anchor sample x_i^s. The indexes of these negative samples are defined as $\mathcal{N}(i) := \{j \in \mathcal{I} \mid y_j \neq y_i,\ j \neq i\}$. To pull positive samples closer together and push negative pairs further in $\mathcal{M}_s$, the learning goal of $f_\phi(\cdot)$ is to minimize the similarity between positive samples and maximize the similarity between positive and negative samples. To optimize $f_\phi(\cdot)$ close to the goal, the contrastive loss of the coarse-grained version is defined as:

$$\mathcal{L}_c = \frac{-1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathcal{L}_c^i, \tag{3}$$

$$\mathcal{L}_c^i = \frac{-1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(s(v_i^s, v_j^s)/\tau)}{\sum_{k \in \mathcal{I} \setminus i} \exp(s(v_i^s, v_k^s)/\tau)}. \tag{4}$$

where $|\cdot|$ denotes the element number of the given set, $\tau \in \mathbb{R}^+$ is a temperature scaling parameter.

3.2.3. *Fine-Grained Version*

The fine-grained version in MGCL was conducted around the same time as the semi-supervised learning method LCLPL (Chaitanya et al., 2023) and had similar inspirations, which fully show the importance of learning local semantic relations. The difference is that we design a lesion-specific fine-grained contrastive strategy, in which representation grains, contrastive classes, and sampling way focus more on learning lesion-related local representations. As described in stage 1 of Fig.2, utilizing patch-level supervisory signal set $\mathcal{D}_p$, the fine-grained CL version aims to make the models gain the local prior representations by a learnable decoder way. To make models focus on the local regions, the weights of the encoder $f_\phi(\cdot)$ learned from the coarse-grained version are first frozen to reserve the global perception ability for the ischemic lesions on the slice level. In the fine-grained version, the patches are augmented online by the manner of augmenting the given single slice, which improves the computational efficiency of the GPU. An augmented slice x_i^s is projected to latent space $\mathcal{M}_p$ as n^2 $L2$-normalized d-dimension patch-level embedding $z_{u,v}^p$. $z_{u,v}^p$ is embedding at the u-th column and v-th row of the decoder output. Thus, the given slice-level embedding $\mathcal{Z}^p := \{z_{u,v}^p | 0 \leq u, v \leq n\}$ including n^2 patch-level embedding $z_{u,v}^p$ is calculated as:

$$\mathcal{Z}^p = H_p(f_\theta(f_\phi(A_p(x_i^s)))), \tag{5}$$

where $A_p(\cdot)$ is the online data augmentation of coarse-grained version, $\mathrm{f}_\theta(\cdot)$ is the learnable decoder, and $H_p(\cdot)$ is the projection head in the fine-grained version. The similarity between the given two patch embeddings $z_{u,v}^p$ and $z_{\hat{u},\hat{v}}^p$ is measured as.

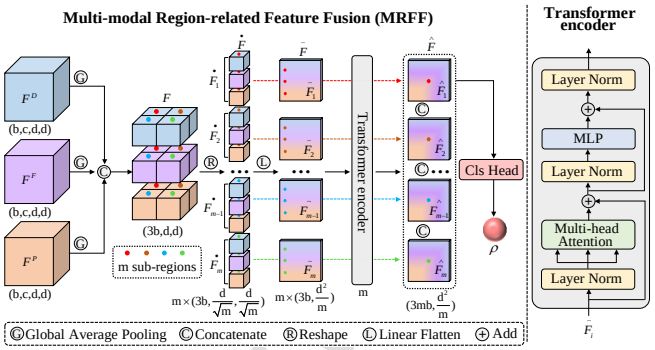

Fig. 3: The overall pipeline of the proposed Multi-modal Region-related Feature Fusion module.

$$s(z_{u,v}^p, z_{\hat{u},\hat{v}}^p) = \frac{(z_{u,v}^p)^{\mathrm{T}} z_{\hat{u},\hat{v}}^p}{\|z_{u,v}^p\| \|z_{\hat{u},\hat{v}}^p\|}, \tag{6}$$

Given a randomly sampled batch, Ω denotes the patch indexes. Set $x_{u,v}^p$ as the anchor sample, $\mathcal{Q}(u,v) := \{\hat{u}, \hat{v} \in \Omega \mid y_{\hat{u},\hat{v}}^p = y_{u,v}^p,\ \hat{u} \neq u \ \vee \ \hat{v} \neq v\}$ represents the set of indexes for all augmented positive samples that are with the same label as $y_{u,v}^p$. The negative samples are these augmented patches with different labels to anchor sample $x_{u,v}^p$. To optimize $f_\theta(\cdot)$ to pull patches of the same category together and push the different apart in latent space $\mathcal{M}_p$, the contrastive loss of the fine-grained version is defined as:

$$\mathcal{L}_f = \frac{-1}{|\Omega|} \sum_{(u,v) \in \Omega} \mathcal{L}_f^{u,v}, \tag{7}$$

$$\mathcal{L}_f^{u,v} = \frac{-1}{|\mathcal{Q}(u,v)|} \sum_{(\hat{u},\hat{v}) \in \mathcal{Q}(u,v)} \log \frac{\exp(s(z_{u,v}^p, z_{\hat{u},\hat{v}}^p)/\tau)}{\sum_{(\tilde{u},\tilde{v}) \in \Omega \setminus (u,v)} \exp(s(z_{u,v}^p, z_{\tilde{u},\tilde{v}}^p)/\tau)}. \tag{8}$$

As illustrated in Fig.2, the encoder $f_\phi(\cdot)$ and the decoder $f_\theta(\cdot)$ were respectively pre-trained by different grained versions via the slice-level and patch-level supervisory signals. The learnable encoder $f_\phi(\cdot)$ updates the weights to incentivize the model to learn the global prior representations by optimizing $\mathcal{L}_c$, which enhances the location ability for the ischemic lesions while perceiving the healthy surroundings. Then, the learnable encoder $f_\theta(\cdot)$ updates the weights to incentivize the model to learn the local prior representations by optimizing $\mathcal{L}_f$, which enhances the perception ability for semantic relation between the ischemic lesion regions and other health regions. In stage 2, the pre-trained encoder $f_\phi(\cdot)$ and decoder $f_\theta(\cdot)$ of every MRI imaging sequence are inherited into the designed multi-task learning architecture to transfer task-specific prior knowledge into the ischemic lesion segmentation (Task 1) and TSS classification (Task 2) tasks, which will boost efficient utilization of limited paired MRI data.

3.3. *Multi-Modal Region-Related Feature Fusion*

Inspired by the idea of Transformer(Vaswani et al., 2017), a multi-modal region-related feature fusion (MRFF) module is

designed, which also includes the multi-head self-attention. As shown in Fig. 3, MRFF can capture the feature correlations among corresponding image regions of the paired multi-modal MRI sequences with sub-region as the basic unit, which reduces the undesirable effects from feature correlation calculation of non-corresponding regions to improve generalization ability. Therefore, it can further facilitate TSS decision-making by refining the process of multi-modal feature fusion.

The encoders $f_\phi(\cdot)^D$, $f_\phi(\cdot)^F$, and $f_\phi(\cdot)^P$ first extract deep image features F^D, F^F, and F^P from paired volumetric MRI sequences $z^D, z^F, z^P \in \mathbb{R}^{b\times h\times w}$ (DWI, FLAIR and PWI), where b, h, and w are the slice number, height, and width of the input MRI sequence, respectively. F^D, F^F, and F^P are calculated as:

$$F^D = f_\phi^D(A_f(z^D)), F^F = f_\phi^F(A_f(z^F)), F^P = f_\phi^P(A_f(z^P)) , \quad (9)$$

where $A_f(\cdot)$ is the online data augmentation of TSS classification. Then, the deep image features F^D, F^F, and $F^P \in \mathbb{R}^{b\times c\times d\times d}$ are mapped into a unified feature space and are concatenated on the channel direction, where c and d are the channel number and the size of the extracted feature maps. This process is defined as:

$$F = cat(g(F^D), g(F^F), g(F^P)) , \quad (10)$$

where $g(\cdot)$ is the global average pooling (GAP) function and $cat(\cdot)$ is the concatenation operation. $F\in \mathbb{R}^{3b\times d\times d}$ is the shallow fusion feature not containing multi-modal feature correlation. Then, F is reshaped and divided into $\left\{\dot{F}_i \in \mathbb{R}^{3b\times \frac{d}{\sqrt{m}}\times \frac{d}{\sqrt{m}}}\right\}_{i=1}^{m}$, where m is the number of divided sub-regions with the same size. Intuitively, each $\dot{F}_i$ represents the shallow multi-modal features at given each sub-region. To obtain the low-dimensional feature representation, $\left\{\dot{F}_i\right\}_{i=1}^{m}$ is flatten linearly to be $\left\{\tilde{F}_i \in \mathbb{R}^{3b\times \frac{d^2}{m}}\right\}_{i=1}^{m}$, which is defined as:

$$\left\{\tilde{F}_i = Flatten(\dot{F}_i)\right\}_{i=1}^{m} , \quad (11)$$

Subsequently, m $\tilde{F}_i$ are fed into m transformer encoders with the multi-head self-attention to capture the feature correlations among multi-modal MRI in each sub-region. Especially, $\tilde{F}_i$ is first send to k parallel heads of multi-head self-attention and then is transformed to query $Q_{i,j} \in \mathbb{R}^{3b\times d_k}$, key $K_{i,j} \in \mathbb{R}^{3b\times d_k}$ and value $V_{i,j} \in \mathbb{R}^{3b\times d_v}$ by using three learnable projection matrices (i.e., $W_{i,j}^Q \in \mathbb{R}^{d_k\times \frac{d^2}{m}}$, $W_{i,j}^K \in \mathbb{R}^{d_k\times \frac{d^2}{m}}$, and $W_{i,j}^V \in \mathbb{R}^{d_v\times \frac{d^2}{m}}$) in each head j, where d_k and d_v are the feature dimensions of projection matrices ($d_k = d_v = d/2$). Then, the outputs of the self-attention from all k heads are concatenated to generate the multi-modal fusion feature $\hat{F}_i$ for the corresponding sub-region, which is formulated as:

$$A(K_{i,j}, Q_{i,j}, V_{i,j}) = softmax(\frac{Q_{i,j}K_i^{\mathrm{T}}}{\sqrt{d_k}})V_{i,j} , \quad (12)$$

$$h_{i,j} = A(K_{i,j}, Q_{i,j}, V_{i,j}) , \quad (13)$$

$$\hat{F}_i = cat(h_{1,j}, ..., h_{m,j})W_i^O + \tilde{F}_i , \quad (14)$$

where $\sqrt{d_k}$ is the scaling factor, and $W_i^O \in \mathbb{R}^{md_v\times \frac{d^2}{m}}$ are the projection matrice. Besides, the residual connection is employed to avoid the vanishing gradient problem during the training phase. Then, all m $\hat{F}_i$ are combined to obtain the final multi-modal fusion feature $\hat{F} \in \mathbb{R}^{3mb\times \frac{d^2}{m}}$, which captures the correlations among of multi-modal MRI in each independent sub-region.

$$\hat{F} = cat(\left\{\hat{F}_i\right\}_{i=1}^{m}) , \quad (15)$$

Finally, $\hat{F}$ is fed into a classification head (Cls Head in Fig. 3) to achieve precise TSS decision-making, which is defined as:

$$\rho = \sigma(W^R(ln(g(\hat{F})))) , \quad (16)$$

where $ln(\cdot)$ is the layernorm operation, $W^R \in \mathbb{R}^{1\times 3mb}$ are the weights of the fully connection layer, and $\sigma(\cdot)$ is the sigmoid activation function. $\rho \in [0, 1]$ is the predicted probability for TSS < 4.5h.

4. Experiment Setup

Multi-modal MRI datasets for AIS analysis are very rare. To support this study, we constructed a standard multi-center MRI acute ischemic stroke dataset (MMIS) via data acquisition and processing. The study (data acquisition and processing) was approved by the Medical Ethical Committee of involved hospitals and was adherent to the tenets of the Declaration of Helsinki.

4.1. Datasets

*4.1.1. **Data acquisition***

To construct the dataset that meets clinical criteria, the included patients met multiple inclusion criteria: (a) The AIS patients are within 24 hours of clear symptom onset. (b) The volume of the ischemic lesions needs to be greater than 1 cc (c) The time of the stroke symptom onset and MRI imaging on admission are recorded (d) The images with severe artifacts were excluded. MRI imaging data of all AIS patients meeting the above criteria were retrospectively collected from several stroke centers in China (Nanjing First Hospital, Nanjing, and Affiliated Jiangning Hospital of Nanjing Medical University, Nanjing) during 2016- 2022 in this study. These data are acquired on the Philips echo planar scanner (Ingenia: 3.0-Tesla, 8-channel receiver array head coil). MMIS currently includes three MRI imaging modals: DWI, FLAIR and PWI. The pixel dimension of DWI images varies from 0.893×0.893×6.6 to 1.198×1.198×7.3 mm^3. The pixel dimension of FLAIR images varies from 0.411×0.411×7 to 0.599×0.599×7.3 mm^3. The pixel dimension of PWI images varies from 1.75×1.75×4 to 1.75×1.75×6 mm^3. DWI, FLAIR and PWI images have 18, 18 and 21 slices respectively to cover the cerebrum from top to bottom. As shown in Table 1, the characteristics distribution of enrolled AIS patients is diverse, which supports that the study about MMIS can be as close to the real clinical diagnosis scenario as possible.

MMIS consists of the labeled part M^l and an unlabeled part M^u. The labeled part M^l includes paired multi-modal MRI sequences, which contain pixel-level annotations for ischemic lesions on each MRI modal and patient-level TSS classification labels (ie., TSS < 4.5h or TSS ≥ 4.5h). The unlabeled part

Table 1: Statistical characteristics distribution of AIS patients on MMIS.

Characteristics	Unlabeled Part	Labeled Part
Demographic		
Patients	425	327
Male	261	201
Female	164	126
Age	65.1 ± 18.7	63.7 ± 16.5
Clinical indicators		
TSS	-	7.26 ± 9.24
NIHSS	-	9.12 ± 7.36
Lesion location		
left	202	155
right	225	172
TSS label		
positive (<4.5 hrs)	-	209
negative (≥ 4.5hrs)	-	118

Table 2: The detailed statistics of labeled and unlabeled part on MMIS. (v / s) means volumes / slices.

Modal	Unlabeled Part (v / s)	Labeled Part (v / s)	Total (v / s)
DWI	364 / 6552	327 / 5886	691 / 12438
FLAIR	291 / 5238	327 / 5886	618 / 11124
PWI	366 / 9150	327 / 5886	693 / 15036

M^u includes massive unpaired MRI data, which is without any expert annotations and is utilized to learn prior representations. M^l includes three sub-parts M^l_d, M^l_p, M^l_f, which represent three MRI sequences respectively. Similarly, M^u includes three sub-parts M^u_d, M^u_p,M^u_f. The detailed statistics of the MMIS are listed in Table 2.

4.1.2. ***Data processing***

For M_l, Elastix tool (Klein et al., 2009) firstly is employed to perform rigid registration between DWI in M^l_d, PWI in M^l_p, and FLAIR in M^l_f, and each voxel in the DWI and PWI images was made to correspond to the same anatomical position in FLAIR image. Then, the DWI, FLAIR and PWI images are respaced to 1×1×6 mm^3 and are resampled 512×512×18. Based on processed images, pixel-level semantic labels for ischemic lesions and patient-level labels for TSS classification can be annotated respectively. Especially, Ischemic lesions were manually annotated on each MRI modal using 3D slicer software (Fedorov et al., 2012), and this process was performed strictly by three radiologists with beyond 6-year of clinical experience. Besides, a senior imaging expert with 15-years experience performed annotation quality control. TSS of each patient was calculated by subtracting the time at which the stroke symptoms were first observed from the time at which the first MRI imaging was obtained. Then, TSS is binarized into two classes: positive (<4.5 hrs) and negative (≥ 4.5hrs). Finally, the patient-level TSS classification labels can be obtained. For M^u, all images are resampled 512×512×N(18/21) and standardized via the z-score way to meet the input requirement of MGCL training.

4.2. Implementation Details

The proposed MGCL is implemented by PyTorch 1.8.0 (Paszke et al., 2019) and runs on NVIDIA GeForce RTX 3090 GPUs with 24 GB memory. The segmentation network including encoder $f_\phi(\cdot)$ and decoder $f_\theta(\cdot)$ follows the architecture of the u-net network (Ronneberger et al., 2015). Before stage 1, the pseudo-label quality refinement strategy based on self-training is employed to help obtain a higher-quality pseudo-label. To generate the pseudo-labels for each sub-part M^u_d, M^u_p,M^u_f in the unlabeled part M^u, the iterative steps are conducted as follows: (1) First, a segmentation network is trained for 80 epochs on the labeled part M^l. (2) Then, the pseudo labels are generated for the unlabeled part M^u (3) Next, the segmentation network is retrained on the combination of M^u with the pseudo labels of high confidence samples and M^l with ground truth to jointly train the segmentation model for 120 generations. (4) Finally, the higher-quality pseudo labels are re-generated for M^u. Especially, the above process only iterates once, and the confidence threshold for pseudo labels is set to 0.7. In each stage, each network employs Adam optimizer (Kingma and Ba, 2014) for parameter optimization, in which the learning rate, moment, and weight decay are set to $1 \times e^{-4}$, 0.9, $1 \times e^{-5}$, 0.1, respectively. In stage 1, the parameters of the encoder $f_\phi(\cdot)$ are updated by optimizing $\mathcal{L}_c$, in which the batch size is set to 70 and training epoch is set to 120. Then, the parameters of the decoder $f_\theta(\cdot)$ are updated by optimizing $\mathcal{L}_f$, in which the batch size is 6 and training epoch is 200. In the fine-grained CL version, the patch-level supervisory signal set is generated based on an area-ratio threshold t (the area ratio of corresponding lesion mask and patch), which aims to reduce label errors caused by the false positive noise of lesion masks. Therefore, these patches still are considered negative samples when the area ratio is below the threshold t, and otherwise, it is a positive sample. The threshold t and patch size are set to 0.1 and 16×16 (refer to Section 5.7). Besides, the temperature scaling parameter τ is 0.1 in the two CL versions. In stage 2, the entire network including the encoder $f_\phi(\cdot)$ and the decoder $f_\theta(\cdot)$ is fine-tuned to achieve task 1 and task 2 by the multi-task learning mode. In task 1, the parameters of encoder $f_\phi(\cdot)$ and the decoder $f_\theta(\cdot)$ are updated by optimizing the hybrid loss including Dice Loss (Milletari et al., 2016) and Cross-Entropy Loss (weighted ratio 1:1), in which the learning rate is $3 \times e^{-4}$, and the batch size is 16. In task 2, the parameters of the encoder $f_\phi(\cdot)$ are first frozen, and then the parameters of the MRFF module are updated by optimizing Focal Loss (Lin et al., 2017), in which the learning rate is $1 \times e^{-3}$, 24, and the batch size is 24. Especially, online data augmentation including random flipping, rotating, and zooming was utilized to alleviate overfitting in each stage.

4.3. Comparison Settings

The proposed framework is compared to task-specific deep-learning algorithms respectively due to the absence of established methods to simultaneously segment ischemic lesions and classify TSS. The best results of all compared methods were retained to achieve the performance after enough parameter adjustment experiments.

4.3.1. Ischemic lesion segmentation task

To illustrate the superiority of the proposed MGCL for the ischemic lesion segmentation task, three different types of meth-

ods are compared. For fairness, these methods are both implemented based on the u-net network (Ronneberger et al., 2015). Besides, all compared methods were pre-trained on M^u, and are fine-tuned on M^l to achieve ischemic lesion segmentation.

i. **Supervised Learning (BaseLine)**. The parameters of the encoder and decoder were randomly initialized. During the model fine-tuning process, extensive online data augmentation was performed, which yielded a strong baseline.

ii. **Self-Supervised Contrastive Learning** Four well-known contrastive learning methods (SimCLR (Chen et al., 2020), SimSiam (Chen and He, 2021), BYOL (Grill et al., 2020), PCL (Zeng et al., 2021), GCL (Chaitanya et al., 2020)) are employed to pre-train and fine-tune the whole network.

iii. **Semi-supervised Learning** A classic (Mixup (Zhang et al., 2017)) and two state-of-the-art (Semi-CL (Hu et al., 2021), LCLPL (Chaitanya et al., 2023)) semi-supervised learning methods are employed to pre-train the encoder and decoder respectively. Then, the whole network was fine-tuned.

4.3.2. TSS classification task

To illustrate the superiority of the proposed MGCL for the TSS classification task, the proposed method is compared with DWI-FLAIR mismatch, Radiomics features + SVM, 3D ResNet, 3D DenseNet, 3D ResNet-mask, and 3D DenseNet-mask. The DWI-FLAIR mismatch represents the classification way of experienced radiologists using the DWI-FLAIR mismatch model. Radiomics features + SVM employs Pyradiomics open-source library (Van Griethuysen et al., 2017) to extract radionics features and then utilizes support vector machine (SVM) (Burges, 1998) for classification. ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) are universal classification DL-based methods consisting of multiple convolutional layers and skip connections. Especially, because the classification task branch of the proposed method shares the feature extraction backbone with the segmentation task branch, the classification task also indirectly utilizes the annotation information of lesions. For fairness, the ischemic lesion masks multiplied with original MR images are fed into 3D ResNet (ie., 3D ResNet-mask) and 3D DenseNet (ie., 3D DenseNet-mask) to classify TSS.

4.4. Evaluation Strategy and Metrics

To objectively evaluate the performance between different methods on Task 1 and Task 2, fivefold cross-validation is adopted. Especially, the unlabeled part M^u is only used for upstream self-supervised representation learning. The labeled part M^l is utilized for the performance evaluation of the downstream Task 1 and Task 2. M^l is divided into five folds equally. One of the folds is used for testing, and the remaining four folds are used for training. In Table 3, the partial datasets including 10% or 50% data respectively are obtained by randomly sampling the corresponding percentage on each training set of five-fold cross-validation. We repeated the experiment 5 times until all folds were used as the testing set. The final result of each method is the average of five predictions.

To demonstrate the advantages of the proposed method, comparative experiments and ablation studies are performed via quantitative metrics. The performance of ischemic lesion segmentation is evaluated by Dice similarity coefficient (DSC). DSC calculates the similarity of foreground regions in the two MRI images according to Eq.13:

$$DSC(R_P, R_G) = \frac{2|R_P \cap R_G|}{|R_P| + |R_G|} \tag{17}$$

where R_P represents the segmentation region of the predicted result and R_G represents the ground truth region of the ischemic lesion.

The performance of TSS classification is evaluated by accuracy (ACC), sensitivity (SEN), and specificity (SPE). The aforementioned metrics are calculated by Eq.14-16:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{18}$$

$$SEN = \frac{TP}{TP + FN} \tag{19}$$

$$SPE = \frac{TN}{TN + FP} \tag{20}$$

where TP, TN, FP, and FN were regarded as true positive, true negative, false positive, and false negative values, respectively. Besides, the area under the receiver operating characteristic curve (AUROC) is also calculated to evaluate the TSS classification ability of the different models. In all quantitative experiments, the higher metrics mean better performance. Besides, the paired t-test is applied in statistical significant analysis.

5. Experiment Results

5.1. Segmentation Comparison With Limited Labels

Table 3 depicts the segmentation results of the comparative study on M^l_d, M^l_p, and M^l_f respectively, and the best results are highlighted with boldface. The training data with different sizes (10%, 50%, and 100%) is utilized to comprehensively compare the segmentation performance of different methods. The final results of different methods are presented in the way of the average scores and standard deviations. Firstly, Baseline means the trained U-net network using randomly initialized parameters and strong online data augmentation. These universal contrastive learning methods including SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), and SimSiam (Chen and He, 2021) are only better than the Baseline. This indicates that the commonly applied contrastive learning settings are inefficient for ischemic lesion segmentation because they only learned slice-level representations and unmined medical images of data characteristics. PCL (Zeng et al., 2021) obtained some performance improvements, which reasonably leveraged the position information in volumetric medical images to generate contrastive sample pairs. GCL (Chaitanya et al., 2020) achieved more competitive performance because it additionally learned local representation. Compared to GCL (Chaitanya et al., 2020), Semi-CL (Hu et al., 2021) can achieve better segmentation because it additionally learns local representations. Similarly, LCLPL (Chaitanya et al., 2023) also learns a semantic-guided local representation. For the training setting with different sizes (10%, 50%, and 100%), our MGCL

Table 3: Performance comparison (Dice) between other methods and our MGCL with limited training data on three datasets (M_d^l, M_p^l and M_f^l) (Mean ± Standard Deviation). Best results are marked in bold.

Method	M_d^l			M_p^l			M_f^l		
	10% data	50% data	100% data	10% data	50% data	100% data	10% data	50% data	100% data
Supervised Learning (BaseLine)									
Random init	0.614±0.053	0.735±0.048	0.778±0.039	0.636±0.041	0.759±0.037	0.795±0.033	0.375±0.111	0.484±0.082	0.646±0.069
Self-Supervised Contrastive Learning									
SimCLR (Chen et al., 2020)	0.621±0.045	0.729±0.042	0.771±0.036	0.640±0.038	0.764±0.034	0.798±0.028	0.373±0.105	0.485±0.081	0.648±0.066
BYOL (Grill et al., 2020)	0.637±0.038	0.735±0.034	0.781±0.033	0.667±0.041	0.773±0.032	0.799±0.027	0.381±0.098	0.489±0.071	0.647±0.068
SimSiam (Chen and He, 2021)	0.642±0.049	0.736±0.035	0.777±0.037	0.671±0.036	0.778±0.029	0.796±0.031	0.378±0.106	0.486±0.075	0.649±0.067
PCL (Zeng et al., 2021)	0.654±0.046	0.742±0.037	0.779±0.035	0.687±0.034	0.781±0.028	0.803±0.029	0.402±0.095	0.497±0.067	0.655±0.052
GCL (Chaitanya et al., 2020)	0.672±0.050	0.747±0.039	0.787±0.032	0.689±0.033	0.783±0.031	0.811±0.028	0.413±0.081	0.508±0.068	0.658±0.062
Semi-supervised Learning									
Mixup (Zhang et al., 2017)	0.665±0.034	0.743±0.035	0.771±0.031	0.681±0.036	0.777±0.026	0.806±0.024	0.397±0.078	0.492±0.080	0.651±0.067
Semi-CL (Hu et al., 2021)	0.681±0.043	0.753±0.036	0.794±0.033	0.702±0.032	0.789±0.029	0.815±0.027	**0.433±0.085**	0.518±0.072	0.664±0.051
LCLPL (Chaitanya et al., 2023)	0.694±0.045	0.761±0.039	0.785±0.034	0.721±0.039	0.788±0.024	0.813±0.022	0.424±0.089	0.505±0.069	0.659±0.069
MGCL(ours)	**0.729±0.036**	**0.776±0.033**	**0.809±0.032**	**0.753±0.035**	**0.803±0.025**	**0.826±0.026**	0.431±0.084	**0.532±0.066**	**0.667±0.054**

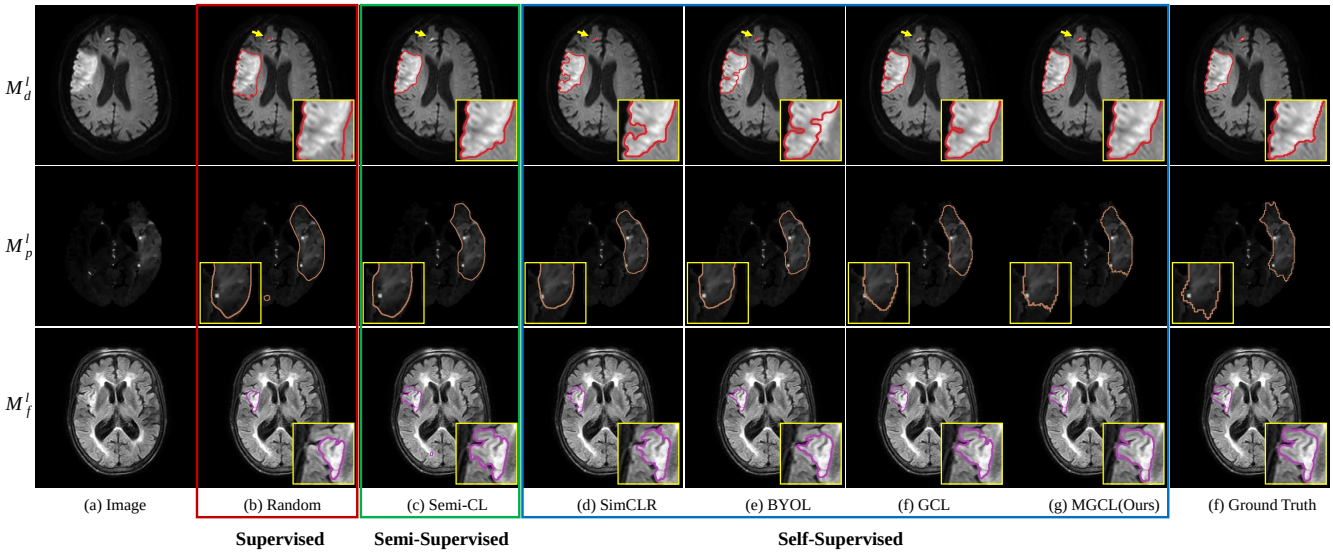

Fig. 4: Visualized comparison of segmentation results on M_d^l, M_p^l and M_f^l with 50% training data. (a) and (b) represent input slice and ground truth. (b) (g) represent the segmentation results of other methods and our MGCL.

observably outperforms BaseLine by 11.5%, 5.1% and 2.1% Dice on M_d^l, 11.7%, 4.4% and 3.1% Dice on M_p^l, 5.6%, 4.8% and 2.1% Dice on M_f^l, respectively. Besides, MGCL achieved the best performance among the above contrastive/semi-supervised learning methods in almost all settings and this may be because of the followings: (a) These methods ((Chen et al., 2020; Grill et al., 2020; Chen and He, 2021)) only learn global representations and it is usually not sufficient for dense prediction tasks. (b) They fully consider domain knowledge to construct contrastive sample pairs for anatomical structure segmentation, but these methods (((Zeng et al., 2021; Chaitanya et al., 2020))) usually cannot be transferred naively when there are significant domain differences. (c) Differences in task scenarios lead to differences in performance. These latest methods ((Hu et al., 2021; Chaitanya et al., 2023)) are designed for very limited labeled settings, such as 2 and 8 samples. For our MGCL, two interesting observations can be found in Table 3: 1) Performance gains is more significant with less training data. This is because the model fully learns prior representations that are closely related to the ischemic lesion segmentation problem in advance. The performance gains become lesser be saturated when the number of training samples gradually increases. This is because with more training samples, the information difference between the training set for fine-tuning and the training set for self-supervised learning becomes small and the fine-tuning performance saturates. 2) On M_d^l and M_p^l, ours with 50% training data can approach or surpass the Baseline performance with 100% training data, which demonstrates that the learned prior representations from MGCL can efficiently mine label information. Thus, the above observations show that our method can significantly alleviate model dependence on training data.

Fig. 4 visualizes the segmentation results of the different methods. Each method is fine-tuned on M_d^l(the first row), M_p^l(the second row), and M_f^l(the third row) with 50% training data. (a) is the input image of three MRI modals, (b)~(g) are the predicted segmentation results of the other methods and ours, and (f) are

Table 4: Performance comparison of different methods for TSS classification task (Mean ± Standard Deviation). *P*-values of the proposed MGCL vs. other methods are indicated by * (<0.05) and + (>0.05). Best results are marked in bold.

Method	Classification			
	ACC	SEN	SPE	AUROC
DWI-FLAIR mismatch	0.688±0.038*	0.629±0.034*	0.759±0.018*	0.719±0.037*
Radiomics features + SVM (Van Griethuysen et al., 2017)	0.703±0.061*	0.639±0.055*	0.786±0.030*	0.741±0.035*
ResNet (He et al., 2016)	0.721±0.015*	0.763±0.020*	0.661±0.024*	0.799±0.019*
ResNet-mask	0.765±0.045*	0.722±0.046*	0.821±0.046*	0.818±0.066+
DenseNet (Huang et al., 2017)	0.734±0.054*	**0.829±0.054**+	0.621±0.022*	0.803±0.011*
DenseNet-mask	0.797±0.047+	0.759±0.035*	0.806±0.021*	0.842±0.031+
Ours	**0.844±0.029**	0.824±0.032	**0.867±0.025**	**0.861±0.028**

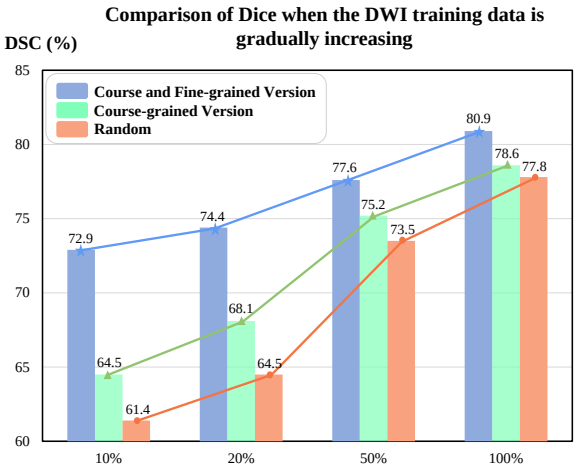

Fig. 5: As the number of DWI training data decreases, ablation studies of the different grained versions in our MGCL are conducted to show the performance superiority on the lesion segmentation task.

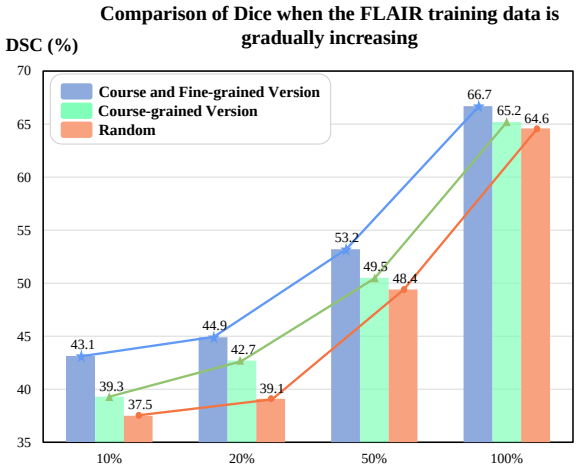

Fig. 6: As the number of FLAIR training data decreases, ablation studies of the different grained versions in our MGCL are conducted to show the performance superiority on the lesion segmentation task.

the ground truth annotations. It can be easily seen that the visualized results predicted by MGCL are more consistent with ischemic lesion boundaries of ground truth than other methods, which are also consistent with the quantitative results in Table 3. Especially, the visual comparison also shows that MGCL is also quite effective in locating boundaries of ischemic lesions.

5.2. Classification Comparison With Other Methods

Table 4 lists the classification performance of the different methods. Experimental results demonstrate that our MGCL performs other methods in the TSS classification task. The proposed MGCL yields 0.844 accuracy, 0.824 sensitivity, 0.867 specificity, and 0.861 auroc. Moreover, the analysis of statistical significance for the p-values shows that the performance difference between the proposed MGCL and the other six advanced methods is significant. The radiologist-derived DWI-FLAIR mismatch model obtains 0.688 accuracy, 0.629 sensitivity, and 0.759 specificity. It can be seen that almost all methods based on machine features outperform the DWI-FLAIR mismatch model from Table 4. This indicates that efficient utilization of image features is important in TSS classification. Radiomics features (Van Griethuysen et al., 2017) + SVM (Burges, 1998) method performs multi-radiomics feature extraction on multi-modal MRI images, and then SVM performs TSS classification based on the extracted multi-radiomics features. The radiomics-based method achieved 0.703 accuracy, 0.639 sensitivity, and 0.786 specificity, but the performance was less than satisfactory. Compared to the radiomics-based method, ResNet (He et al., 2016) and DenseNet (Huang et al., 2017) obtained better performance (0.721 accuracy and 0.734 accuracy) because convolutional neural networks usually can extract richer deep image features. ResNet-mask and DenseNet-mask can achieve modest performance improvement of 4.4% and 6.3% accuracy respectively because the mask information of lesion annotation can help suppress excessive noise and task-unrelated features in the image background, improving the ability to distinguish effective features from the entire image. Although they can remove noise from the background, lacking the contrast feature between the lesion and the healthy surroundings still limits the performance. Compared to the above methods, our method focuses on task-related regions and perceives healthy surroundings, extracting the contrast feature between the lesion and the surroundings, which achieves the most competitive performance.

5.3. Ablation Studies of MGCL

5.3.1. Ablation Studies on lesion segmentation task

Fig.5-7 demonstrates that our innovations bring significant improvements to the lesion segmentation task, which makes our MGCL more competitive. Specifically, the effectiveness of different grained versions in MGCL is demonstrated on M_d^l, M_p^l, and M_f^l. As shown in Fig.5-7, 0%, 50%, 80%, and 90% of training data are removed respectively on each dataset. It can be seen that with the rapid reduction of the number of training

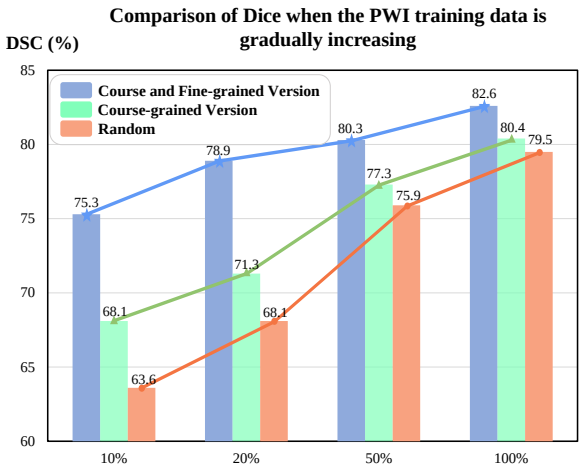

Fig. 7: As the number of PWI training data decreases, ablation studies of the different grained versions in our MGCL are conducted to show the performance superiority on the lesion segmentation task.

Table 5: Ablation studies on TSS classification task.

BackBone	Cls Head	ACC	SEN	SPE	AUROC
MGCL (Random Init)	MLP	0.742	0.806	0.667	0.745
	Transformer	0.763	0.735	0.778	0.779
	MRFF	0.803	0.743	**0.871**	0.831
MGCL (Pre-trained)	MLP	0.761	0.743	0.781	0.763
	Transformer	0.798	0.757	0.828	0.823
	MRFF	**0.844**	**0.824**	0.867	**0.861**

data, the performance of models with different modules starts to decline to different degrees: the random initialized model falls the fastest, the model pre-trained by the coarse-grained CL version is the second, and the model pre-trained by the coarse and fine-grained CL version is the slowest. Compared to the randomly initialized model, MGCL takes the weights pre-trained by the coarse-fined CL version to gain obvious dice improvement of the 11.5% on M_d^l, 11.7% on M_p^l, and 5.6% on M_f^l when the number of training data is reduced to 10%. It is worth noting that our MGCL with only 10% training data can approach the performance of the random initialized model with only 50% training data on M_d^l and M_p^l. Therefore, the experiment results indicate that our innovation including the coarse-grained and the fine-grained CL versions of MGCL can both improve the performance of the ischemic lesion segmentation task.

5.3.2. Ablation Studies on TSS classification Task

Table 5 demonstrates our innovations display significant improvements for the TSS classification task. The backbone of the classification task (Fig.2) is reserved to investigate the influence of each innovation (ie., coarse-grained CL version of MGCL, MRFF). First, the effect of the coarse-grained CL version is demonstrated by comparison with the backbone with random initialized weights. Second, in each backbone with different initialization weights, well-known classifiers are compared to demonstrate the effect of MRFF. With the backbone of the random initialized weights, our MRFF achieves the best performance in all classifiers, which obtains 0.803 accuracy, 0.743 sensitivity, 0.871 specificity, and 0.831 auroc. Compared to MLP or Transformer(Vaswani et al., 2017), MRFF can obtain 6.1% and 4.0% accuracy improvement respectively in two different backbones. Results indicate that MRFF explicitly computes and represents the DWIF-LAIR and DWI-PWI mismatch patterns on the feature level, and fuses this information to extracted deep features finally, which can promote TSS classification performance. With the backbone pre-trained by the coarse-grained CL version, MLP, Transformer(Vaswani et al., 2017), and MRFF achieved 1.9%, 2.5%, and 4.1% accuracy improvements, respectively. Results indicate that the coarse-grained version of MGCL utilizes the learned global prior representations to locate the ischemic lesions while perceiving the healthy surroundings, extract task-related features, and finally contribute to accurate TSS classification.

Table 6: Performance comparison with different MRI modal combination on TSS classification task.

modals			Classification			
DWI	FLAIR	PWI	ACC	SEN	SPE	AUROC
✓			0.738	0.667	0.861	0.766
	✓		0.619	0.614	0.828	0.594
		✓	0.580	0.363	0.768	0.642
✓	✓		0.813	0.804	0.828	0.824
✓		✓	0.769	0.686	**0.867**	0.801
✓	✓	✓	**0.844**	**0.824**	**0.867**	**0.861**

5.4. Comparison with Different Modals Combination

Table 6 shows the TSS classification performance for each combined case of different MRI modals using our proposed method (single or multiple modal branches in Fig.2), which investigates the contribution of different modals for TSS classification. Classification performance is inefficient when predicting TSS using a single DWI, FLAIR, or PWI image. There is a certain performance improvement compared to a single modal when leveraging the combination of two modals. This indicates that the TSS classification performance can be improved when the model incorporates the latent knowledge of DWI-FLAIR or DWI-PWI mismatch patterns using the MRFF module. It achieves the best performance when all modals are available. This indicates that the model incorporates the diagnostic knowledge of multiple mismatch patterns and fuses this knowledge to extracted deep features, which further facilitates TSS classification performance.

5.5. Representation Analysis of Different Grained Versions

The learned prior representations from coarse-grained and fine-grained CL versions of MGCL are visualized on M_d^l, M_p^l and M_f^l without using the annotation information to demonstrate the discrimination ability for different lesion samples on the slice level and patch level. Specifically, the proposed MGCL is firstly trained on M_d^u, M_p^u, and M_f^u. After passing randomly selected data on M_d^l, M_p^l and M_f^l using the trained MGCL, the learned different level representations are obtained from the projection head $H_s(\cdot)$ and $H_p(\cdot)$ respectively. Finally, the t-Distributed Stochastic

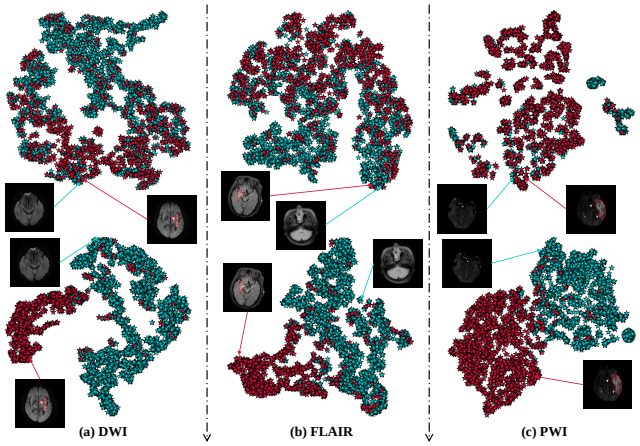

Fig. 8: Visualize the global representations learned from the coarse-grained version of MGCL. (a), (b), (c) mean applying t-SNE for the DWI, FLAIR and PWI slices respectively.

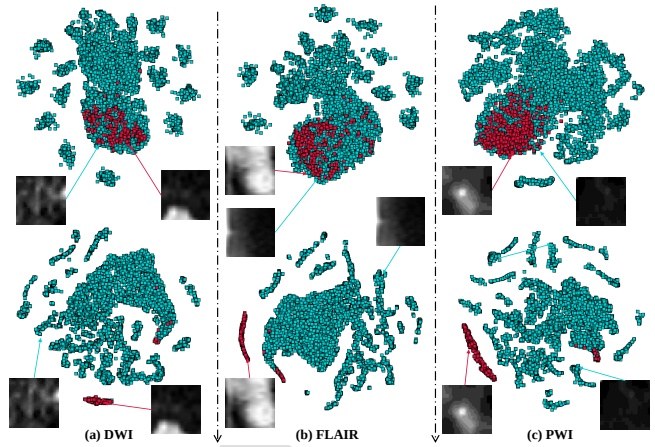

Fig. 9: Visualize the local representations learned from the fine-grained version of MGCL. (a), (b), (c) mean applying t-SNE for the DWI, FLAIR and PWI patches respectively.

Neighbor Embedding (t-SNE) method (Van der Maaten and Hinton, 2008) is employed to visualize the learned representations.

Slice-Level Global Representation. Fig.8 visualizes learned representations after applying t-SNE on slice-level for three MRI modal data. In each column, the first cluster (from top to bottom) is the slice-level representations generated by the randomly initialized model, the second cluster is generated from the model pre-trained by the coarse-grained CL version. It can be easily seen that slice-level representations of different categories (the slice with/without ischemic lesions) generated by the randomly initialized model are mixed in the low latent space, while these slice representations from the coarse-grained CL version are clustered into two meaningful groups. In this cluster, most of the slices in each group belong to the same categories. Besides, it can be seen that the representations of the same class slices are pulled closely (indicated by their coordinates) and the different are pushed apart in low latent space. This indicates that the coarse-grained CL version can facilitate the model to gain the global discriminative ability for lesions.

Patch-Level Local Representation. Fig.9 visualizes learned representations after applying t-SNE on patch-level for three modal data. Fig.9 is consistent with the description structure of Fig.8. The fine-grained CL version can facilitate pushing the patches of different categories (the patch with/without ischemic lesions) apart according to the extracted semantic information, and pull patches belong to the same categories closer together to form different clusters. This indicates that the fine-grained CL version can facilitate the model to gain local perception ability for semantic relation between the ischemic lesion region and other health regions.

Therefore, the learned global and local prior representations can promote easily achieving competitive segmentation and classification performance in limited annotation settings.

5.6. Perception Ability Analysis of Task-Related Regions

Fig. 10 shows that EigenCAM (Muhammad and Yeasin, 2020) is employed to generate the perception maps that analyze the perception for different areas on three MRI modal. As described in Fig. 10 (a), the lesion regions are delineated on three different MRI models by the red circle. For the observation sake, the perception maps are overlapped with the corresponding input slices. l^k represents the perception map from the k-th layer of the encoder. l^{-1} represents the perception map from the last layer of the encoder. In Fig.10 (b), it can be seen that the randomly initialized model can not perceive the lesion regions. Compared to (b), the model trained by MGCL can efficiently locate the ischemic lesions while perceiving the healthy surroundings. With the increment of k (from (c) to (f) in Fig. 10), the perception for the ischemic lesions is gradually enhanced, which can generate accurate perception in the last layer of the encoder. Therefore, the model that transfers the global prior representations can locate ischemic lesions and perceive healthy surroundings, guiding the segmentation and classification tasks to efficiently find task-related regions, and extracting the task-related features, which mitigates overfitting caused by the limited training data.

Table 7: Performance comparison (Dice) at different area-ratio thresholds t on three datasets (M_d^l, M_p^l and M_f^l) (Mean ± Standard Deviation).

t	M_d^l		M_p^l		M_f^l	
	10% data	50% data	10% data	50% data	10% data	50% data
0.05	0.717±0.039	0.766±0.032	**0.759±0.034**	**0.804±0.027**	0.425±0.088	0.528±0.073
0.1	**0.729±0.036**	**0.775±0.033**	0.753±0.035	0.803±0.025	**0.431±0.084**	**0.532±0.066**
0.2	0.713±0.044	0.759±0.037	0.738±0.033	0.793±0.029	0.422±0.069	0.524±0.072
0.3	0.692±0.042	0.751±0.040	0.721±0.037	0.780±0.030	0.411±0.071	0.521±0.068

5.7. Efficiency Analysis of Area-ratio Threshold

To investigate the efficiency of the area-ratio threshold t on the lesion segmentation task, the performance at different thresholds is compared quantitatively on three datasets (M_d^l, M_p^l and M_f^l). As listed in Table 7, the lesion segmentation task achieves optimal performance when threshold t is set to 0.1. When the threshold t is set larger, the performance starts to drop gradually. This fact is attributed to (1) These patches with relatively big lesions also start to be set to negative samples when t increases. The caused label errors are detrimental to the local representation learning in the fine-grained CL version. (2) The area ratio of the lesion mask and patch is relatively small, and these patches

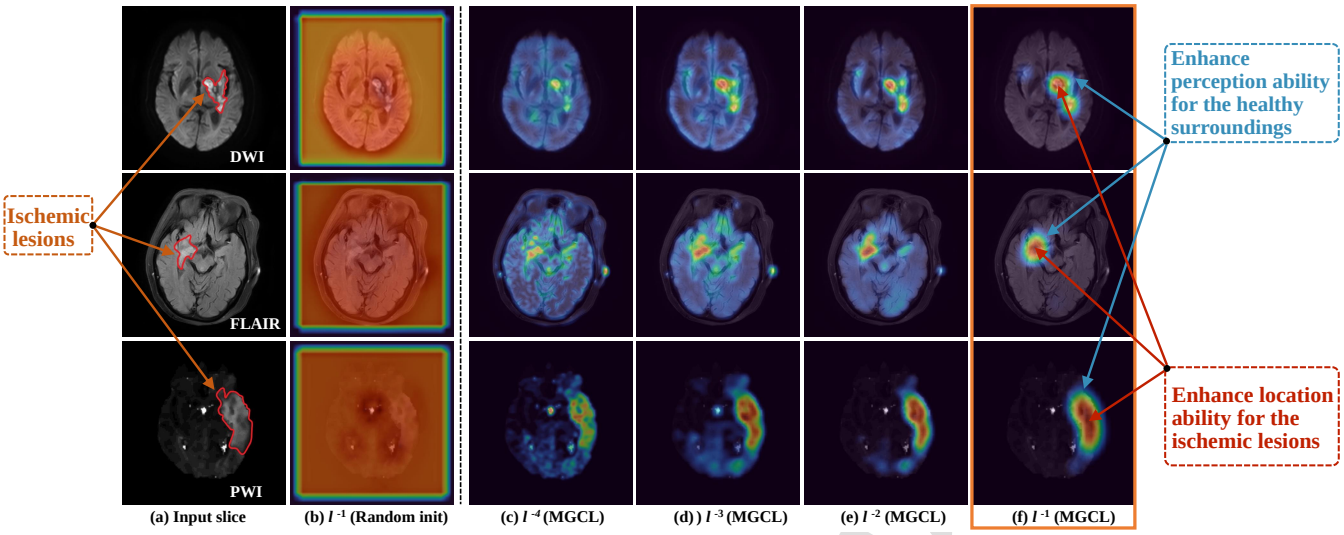

Fig. 10: Perception ability analysis of task-related regions. l^k represents the perception map from the k-th layer of the encoder. With the increment of k, MGCL can progressively perceive the precise task-related regions including ischemic lesion region and healthy surroundings.

usually contain rich edge information. Therefore, when t becomes larger, these patches are considered as negative samples and then the learning efficiency for edge features may become lower gradually.

6. Discussion and Conclusion

Identifying the unknown TSS for unwitnessed AIS patients from multi-modal MRI imaging is a challenging, but clinically meaningful task for better stroke evaluation and treatment decision-making. DWI-FLAIR mismatch model is now clinically recommended to classify TSS for guiding tPA thrombolysis. However, the mismatch model may not find all patients within 4.5 hrs because of its simplicity (Odland et al., 2015). Due to the advantages of convolutional neural networks in feature extraction, related deep learning-based methods were developed to improve the TSS classification performance. Limited by the urgency of AIS onset and medical conditions, it is difficult to collect large-scale paired multi-modal MRI imaging sequences. Thus, data-driven methods are easier to learn irrelevant features in the TSS classification because of the lack of sufficient training data, which leads to poor generalization ability. We observed that unpaired MRI data are available, but are less often considered and underexplored. Thus, we proposed a novel multi-grained contrastive learning (MGCL) framework to fully develop large-scale unpaired unlabeled data to incentivize efficient utilization of the limited paired data, which facilitates the performance improvement of the ischemic lesion segmentation and TSS classification tasks. Specifically, MGCL achieves efficient AIS analysis via two cascade stages: Stage 1 encourages the models to learn prior representations from massive unlabeled unpaired data based on two task-specific contrastive learning versions: (a) The coarse-grained CL version encourages the models to learn global prior representations to enhance the location ability for the ischemic lesions while perceiving the healthy surroundings, which helps extract the task-related deep features. (b) The fine-grained CL version encourages the models to learn local prior representations to enhance the discriminative ability for the ischemic lesion regions, which helps supplement the lesion details. Finally, learned global and local prior representations are transferred reasonably into a designed multi-task framework, which comprehensively improves the performance of ischemic lesion segmentation and TSS classification tasks. In this process, the proposed multi-modal region-related feature fusion (MRFF) module explicitly calculates the feature correlations among corresponding image regions of the paired multi-modal MRI sequences, which can explicitly integrate the clinical diagnostic knowledge including DWI-FLAIR and DWI-PWI mismatch patterns into TSS decision-making. Extensive experiments on the large-scale multi-center MRI dataset demonstrate the superiority of the proposed MGCL. Our method can simultaneously segment ischemic lesions and classify TSS in an end-to-end multi-task way and outperforms the conventional radiologist-derived DWI-FLAIR mismatch model. Therefore, it is very promising that it helps better stroke evaluation and treatment decision-making.

Recent works (Srinidhi et al., 2022; Gao et al., 2022; Chaitanya et al., 2019, 2021) provide evidence that integrating self-supervised and semi-supervised learning methods usually further boosts the potential of massive unlabeled data. Therefore, how to efficiently integrate these two learning paradigms will become one of our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Key Project of Research and Development Plan under Grants

2022YFC2401600, 2022YFC2408500, and 2022YFE0116700, and in part by the National Natural Science Foundation of China under Grant T2225025 and 82202128, in part by the Key Research and Development Programs in Jiangsu Province of China under Grant BE2021703 and BE2022768.

References

Araslanov, N., Roth, S., 2021. Self-supervised augmentation consistency for adapting semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15384–15394.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al., 2021. Big self-supervised models advance medical image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3478–3488.

Bang, O. Y., S.J.L.K.S.J.K.G.M.C.C.S.O.B...L.D.S., 2011. Collateral flow predicts response to endovascular therapy for acute ischemic stroke. Stroke 90, 101926.

Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R., et al., 2019. Heart disease and stroke statistics—2019 update: a report from the american heart association. Circulation 139, e56–e528.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2, 121–167.

Campbell, B.C., Ma, H., Ringleb, P.A., Parsons, M.W., Churilov, L., Bendszus, M., Levi, C.R., Hsu, C., Kleinig, T.J., Fatar, M., et al., 2019. Extending thrombolysis to 4· 5–9 h and wake-up stroke using perfusion imaging: a systematic review and meta-analysis of individual patient data. The Lancet 394, 139–147.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2020. Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in Neural Information Processing Systems 33, 12546–12558.

Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E., 2023. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Medical Image Analysis 87, 102792.

Chaitanya, K., Karani, N., Baumgartner, C.F., Becker, A., Donati, O., Konukoglu, E., 2019. Semi-supervised and task-driven data augmentation, in: Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26, Springer. pp. 29–41.

Chaitanya, K., Karani, N., Baumgartner, C.F., Erdil, E., Becker, A., Donati, O., Konukoglu, E., 2021. Semi-supervised task-driven data augmentation for medical image segmentation. Medical Image Analysis 68, 101934.

Chen, C., Wang, Y., Niu, J., Liu, X., Li, Q., Gong, X., 2021. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. IEEE Transactions on Medical Imaging 40, 2439–2451.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, X., He, K., 2021. Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758.

Davis, S.M., Donnan, G.A., Parsons, M.W., Levi, C., Butcher, K.S., Peeters, A., Barber, P.A., Bladin, C., De Silva, D.A., Byrnes, G., et al., 2008. Effects of alteplase beyond 3 h after stroke in the echoplanar imaging thrombolytic evaluation trial (epithet): a placebo-controlled randomised trial. The Lancet Neurology 7, 299–309.

Ebinger, M., Galinovic, I., Rozanski, M., Brunecker, P., Endres, M., Fiebach, J.B., 2010. Fluid-attenuated inversion recovery evolution within 12 hours from stroke onset: a reliable tissue clock? Stroke 41, 250–255.

Emeriau, S., Serre, I., Toubas, O., Pombourcq, F., Oppenheim, C., Pierot, L., 2013. Can diffusion-weighted imaging–fluid-attenuated inversion recovery mismatch (positive diffusion-weighted imaging/negative fluid-attenuated inversion recovery) at 3 tesla identify patients with stroke at¡ 4.5 hours? Stroke 44, 1647–1651.

Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al., 2012. 3d slicer as an image computing platform for the quantitative imaging network. Magnetic resonance imaging 30, 1323–1341.

Galinovic, I., Puig, J., Neeb, L., Guibernau, J., Kemmling, A., Siemonsen, S., Pedraza, S., Cheng, B., Thomalla, G., Fiehler, J., et al., 2014. Visual and region of interest–based inter-rater agreement in the assessment of the diffusion-weighted imaging–fluid-attenuated inversion recovery mismatch. Stroke 45, 1170–1172.

Gao, Z., Jia, C., Li, Y., Zhang, X., Hong, B., Wu, J., Gong, T., Wang, C., Meng, D., Zheng, Y., et al., 2022. Unsupervised representation learning for tissue segmentation in histopathological images: From global to local contrast. IEEE Transactions on Medical Imaging 41, 3611–3623.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284.

Han, Y., Chen, C., Tewfik, A., Glicksberg, B., Ding, Y., Peng, Y., Wang, Z., 2022. Knowledge-augmented contrastive learning for abnormality classification and localization in chest x-rays with radiomics using a feedback loop, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2465–2474.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Ho, K.C., Speier, W., Zhang, H., Scalzo, F., El-Saden, S., Arnold, C.W., 2019. A machine learning approach for classifying ischemic stroke onset time from imaging. IEEE transactions on medical imaging 38, 1666–1676.

Hu, X., Zeng, D., Xu, X., Shi, Y., 2021. Semi-supervised contrastive learning for label-efficient medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 481–490.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Jiang, L., Sun, J., Wang, Y., Yang, H., Chen, Y.C., Peng, M., Zhang, H., Chen, Y., Yin, X., 2024. Diffusion-/perfusion-weighted imaging fusion to automatically identify stroke within 4.5 h. European Radiology , 1–12.

Jiang, L., Wang, S., Ai, Z., Shen, T., Zhang, H., Duan, S., Chen, Y.C., Yin, X., Sun, J., 2022. Development and external validation of a stability machine learning model to identify wake-up stroke onset time from mri. European Radiology 32, 3661–3669.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2009. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 29, 196–205.

Kolesnikov, A., Zhai, X., Beyer, L., 2019. Revisiting self-supervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1920–1929.

Kong, J., He, Y., Zhu, X., Shao, P., Xu, Y., Chen, Y., Coatrieux, J.L., Yang, G., 2022. Bkc-net: Bi-knowledge contrastive learning for renal tumor diagnosis on 3d ct images. Knowledge-Based Systems 252, 109369.

Lee, H., Lee, E.J., Ham, S., Lee, H.B., Lee, J.S., Kwon, S.U., Kim, J.S., Kim, N., Kang, D.W., 2020. Machine learning approach to identify stroke within 4.5 hours. Stroke 51, 860–866.

Li, X., Jia, M., Islam, M.T., Yu, L., Xing, L., 2020. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. IEEE Transactions on Medical Imaging 39, 4023–4033.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., 2021. Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge and Data Engineering .

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.

McLeod, D.D., Parsons, M.W., Levi, C.R., Beautement, S., Buxton, D., Roworth, B., Spratt, N.J., 2011. Establishing a rodent stroke perfusion computed tomography model. International Journal of Stroke 6, 284–289.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE. pp. 565–571.

Misra, I., Maaten, L.v.d., 2020. Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, pp. 6707–6717.
Moradiya, Y., Janjua, N., 2013. Presentation and outcomes of "wake-up strokes" in a large randomized stroke trial: analysis of data from the international stroke trial. Journal of Stroke and Cerebrovascular Diseases 22, e286–e292.
Muhammad, M.B., Yeasin, M., 2020. Eigen-cam: Class activation map using principal components, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–7.
Murphy, B., Chen, X., Lee, T.Y., 2007. Serial changes in ct cerebral blood volume and flow after 4 hours of middle cerebral occlusion in an animal model of embolic cerebral ischemia. American Journal of Neuroradiology 28, 743–749.
Odland, A., Særvoll, P., Advani, R., Kurz, M.W., Kurz, K.D., 2015. Are the current mri criteria using the dwi-flair mismatch concept for selection of patients with wake-up stroke to thrombolysis excluding too many patients? Scandinavian journal of trauma, resuscitation and emergency medicine 23, 1–6.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.
Pedersen, M., Andersen, M.B., Christiansen, H., Azawi, N.H., 2020. Classification of renal tumour using convolutional neural networks to detect oncocytoma. European Journal of Radiology 133, 109343.
Polson, J.S., Zhang, H., Nael, K., Salamon, N., Yoo, B.Y., El-Saden, S., Starkman, S., Kim, N., Kang, D.W., Speier, W., et al., 2022. Deep learning approaches to identify patients within the thrombolytic treatment window. medRxiv .
Powers, W.J., Rabinstein, A.A., Ackerson, T., Adeoye, O.M., Bambakidis, N.C., Becker, K., Biller, J., Brown, M., Demaerschalk, B.M., Hoh, B., et al., 2019. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the american heart association/american stroke association. Stroke 50, e344–e418.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.
Srinidhi, C.L., Kim, S.W., Chen, F.D., Martel, A.L., 2022. Self-supervised driven consistency training for annotation efficient histopathology image analysis. Medical Image Analysis 75, 102256.
Thomalla, G., Cheng, B., Ebinger, M., Hao, Q., Tourdias, T., Wu, O., Kim, J.S., Breuer, L., Singer, O.C., Warach, S., et al., 2011. Dwi-flair mismatch for the identification of patients with acute ischaemic stroke within 4· 5 h of symptom onset (pre-flair): a multicentre observational study. The Lancet Neurology 10, 978–986.
Thomalla, G., Gerloff, C., 2015. Treatment concepts for wake-up stroke and stroke with unknown time of symptom onset. Stroke 46, 2707–2713.
Thomalla, G., Rossbach, P., Rosenkranz, M., Siemonsen, S., Krützelmann, A., Fiehler, J., Gerloff, C., 2009. Negative fluid-attenuated inversion recovery imaging identifies acute ischemic stroke at 3 hours or less. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 65, 724–732.
Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer research 77, e104–e107.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.
Vijayan, M., Reddy, P.H., 2016. Peripheral biomarkers of stroke: focus on circulatory micrornas. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1862, 1984–1993.
Wolman, D.N., Iv, M., Wintermark, M., Zaharchuk, G., Marks, M.P., Do, H.M., Dodd, R.L., Albers, G.W., Lansberg, M.G., Heit, J.J., 2018. Can diffusion-and perfusion-weighted imaging alone accurately triage anterior circulation acute ischemic stroke patients to endovascular therapy? Journal of neurointerventional surgery 10, 1132–1136.
Wu, Y., Zeng, D., Wang, Z., Shi, Y., Hu, J., 2022. Distributed contrastive learning for medical image segmentation. Medical Image Analysis 81, 102564.
Xu, X., Wang, C., Guo, J., Gan, Y., Wang, J., Bai, H., Zhang, L., Li, W., Yi, Z., 2020. Mscs-deepln: Evaluating lung nodule malignancy using multi-scale cost-sensitive neural networks. Medical Image Analysis 65, 101772.
Yang, P., Yin, X., Lu, H., Hu, Z., Zhang, X., Jiang, R., Lv, H., 2022. Cs-co: A hybrid self-supervised visual representation learning method for h&e-stained histopathological images. Medical Image Analysis 81, 102539.
Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A., 2016. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE transactions on medical imaging 36, 994–1004.
Zabihollahy, F., Schieda, N., Krishna, S., Ukwatta, E., 2020. Automated classification of solid renal masses on contrast-enhanced computed tomography images using convolutional neural network with decision fusion. European Radiology 30, 5183–5190.
Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y., 2021. Positional contrastive learning for volumetric medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 221–230.
Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 .
Zhang, H., Polson, J.S., Nael, K., Salamon, N., Yoo, B., El-Saden, S., Scalzo, F., Speier, W., Arnold, C.W., 2021. Intra-domain task-adaptive transfer learning to determine acute ischemic stroke onset time. Computerized Medical Imaging and Graphics 90, 101926.
Zhu, H., Jiang, L., Zhang, H., Luo, L., Chen, Y., Chen, Y., 2021. An automatic machine learning approach for ischemic stroke onset time identification based on dwi and flair imaging. NeuroImage: Clinical 31, 102744.
Ziegler, A., Ebinger, M., Fiebach, J.B., Audebert, H.J., Leistner, S., 2012. Judgment of flair signal change in dwi–flair mismatch determination is a challenge to clinicians. Journal of neurology 259, 971–973.

In general, our contributions include the following:

- We propose a novel multi-grained contrastive learning (MGCL) framework based on two cascade stages for the AIS analysis task.
- We design two task-specific contrastive feature enhancement strategies, which help to enhance the global location ability for the ischemic lesions and the local perception ability for semantic relation respectively.
- We develop a multi-modal region-related feature fusion module to adequately capture the feature relationship between multi-modal MRI images.
- Extensive experiments also show the superiority of the proposed framework on a constructed large-scale multi-center multi-modal MRI dataset.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: