Exploring Feature Fusion from A Contrastive Multi-Modality Learner for Liver Cancer Diagnosis

Yang Fan Chiang*
e94081107@gs.ncku.edu.tw
National Cheng Kung University
Tainan, Taiwan

Pei-Xuan Li*
n28121107@gs.ncku.edu.tw
National Cheng Kung University
Tainan, Taiwan

Ding-You Wu*
n26090774@gs.ncku.edu.tw
National Cheng Kung University
Tainan, Taiwan

Hsun-Ping Hsieh†
hphsieh@mail.ncku.edu.tw
National Cheng Kung University
Tainan, Taiwan

Ching-Chung Ko
kocc0729@gmail.com
Chi-Mei Medical Center
Tainan, Taiwan

ABSTRACT

Self-supervised contrastive learning has achieved promising results in computer vision, and recently it also received attention in the medical domain. In practice, medical data is hard to collect and even harder to annotate, but leveraging multi-modality medical images to make up for small datasets has proved to be helpful. In this work, we focus on mining multi-modality Magnetic Resonance (MR) images to learn multi-modality contrastive representations. We first present multi-modality data augmentation (MDA) to adapt contrastive learning to multi-modality learning. Then, the proposed cross-modality group convolution (CGC) is used for multi-modality features in the downstream fine-tune task. Specifically, in the pre-training stage, considering different behaviors from each MRI modality with the same anatomic structure, yet without designing a handcrafted pretext task, we select two augmented MR images from a patient as a positive pair, and then directly maximize the similarity between positive pairs using Simple Siamese networks. To further exploit multi-modality representation, we combine 3D and 2D group convolution with a channel shuffle operation to efficiently incorporate different modalities of image features. We evaluate our proposed methods on liver MR images collected from a well-known hospital in Taiwan. Experiments show our framework has significantly improved from previous methods.

KEYWORDS

Contrastive learning, Multi-modality learning, MRI image analysis

ACM Reference Format:
Yang Fan Chiang, Pei-Xuan Li, Ding-You Wu, Hsun-Ping Hsieh, and Ching-Chung Ko. 2023. Exploring Feature Fusion from A Contrastive Multi-Modality Learner for Liver Cancer Diagnosis. In *ACM Multimedia Asia 2023 (MMAsia '23), December 6–8, 2023, Tainan, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3595916.3626383

*Authors contributed equally to this research.
†Corresponding authors.

ACM Multimedia Asia'23, 2023, Tainan,Taiwan

ACM ISBN 979-8-4007-0205-1/23/12…$15.00
https://doi.org/10.1145/3595916.3626383

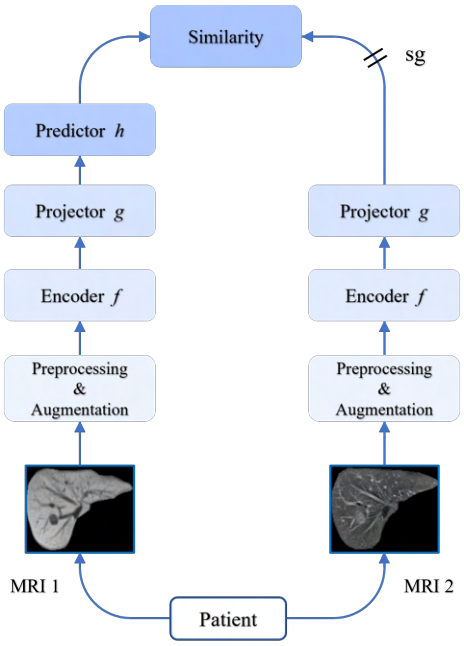

Figure 1: An illustration of self-supervised learning for multi-modality MR images. We randomly select two images from a patient as positive pairs and pass them to the SimSiam networks. sg indicate the stop gradient operation.

1 INTRODUCTION

MRI analysis is crucial for liver cancer diagnosis, where cancer stages are classified into four levels[1]. Traditional MR images can help identify the third or fourth stage visually, but distinguishing the first and second stages is challenging due to subtle and illegible features. Our goal is to propose a machine learning approach that utilizes multi-modality MR images (e.g., T1-weighted or Out-of-phase images) to accurately classify the pathological stage of

[1]https://www.cancer.org/cancer/liver-cancer/detection-diagnosis-staging/staging.html

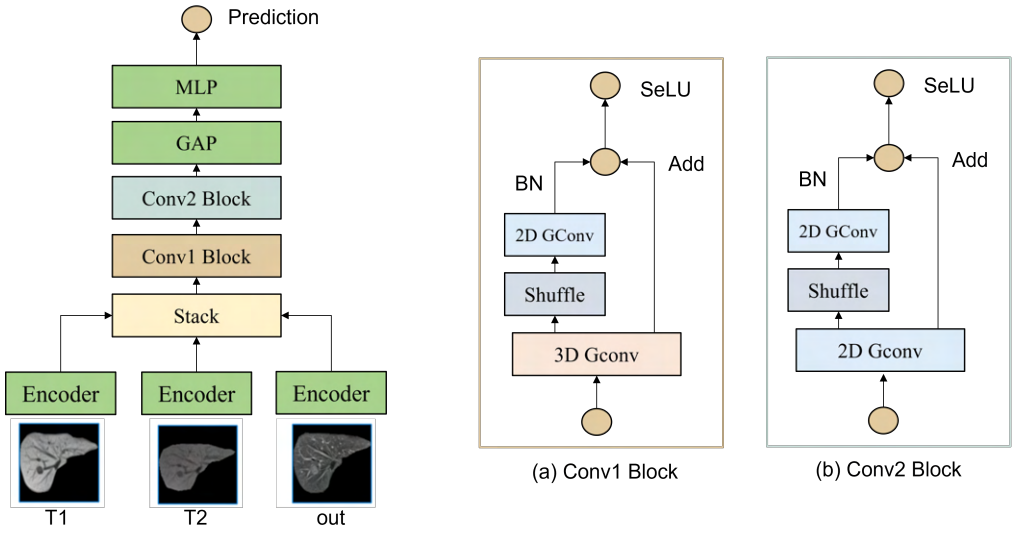

Figure 2: An illustration of the proposed overall architecture. The self-supervised pre-trained model serves as a multi-modality encoder. In the downstream task, we use the CGC layer by combining Conv1 and Conv2 block to incorporate multi-modality features.

liver cancer, assisting doctors in diagnosis. Various modalities of MR images depict the same anatomical structure differently. For instance, edema appears hyperintense in T2-weighted images and hypointense in T1-weighted images. Similarly, abnormal fat accumulation exhibits hypointensity in Out-of-phase imaging. Deep learning has shown promising results in medical tasks such as classification, object detection, and segmentation [1, 5, 8, 18, 19, 22, 25], but acquiring labeled medical data is challenging due to the need for expert inspection and annotation. To address this limitation, we propose using multi-modality MR images as a feasible solution to compensate for inadequate datasets.

Multi-modality learning in the medical domain has been extensively explored. One approach involves jointly training assistant-modality and target-modality images, while another uses transfer learning by pre-training assistant-modality images and fine-tuning target-modality images. However, these methods do not effectively leverage cross-modality information. Previous studies [18] have shown poor performance in joint training and transfer learning due to significant appearance discrepancies, making it challenging to learn cross-modality features directly from multi-modality images.

Self-supervised contrastive learning has gained significant attention in the medical domain [1, 21] for learning effective visual representations [2, 4, 7]. Contrastive learning requires data augmentation to produce different views of training examples, enabling the extraction of effective representations through maximizing the agreement between positive pairs [27]. In this work, our aim is to learn representative multi-modality features from MR images in an unsupervised manner. Inspired by prior works, we propose multi-modality data augmentation (MDA) to adapt contrastive learning for multi-modality learning. By considering MR images as different views of a naive anatomic representation, MDA maximizes agreement between augmented MR images' features, enabling the extraction of effective contrastive representations from multi-modality data.

In this paper, we propose a self-supervised pre-trained multi-modality encoder for the downstream supervised task. Our goal is to enhance model robustness by exploiting representative multi-modality features. Tseng et al. [25] introduced 3D cross-modality convolution (CMC) to leverage different modality features from MR images. However, the high computation cost of 3D convolution is detrimental for small datasets. To address this, we present a light-weight solution called cross-modality group convolution (CGC), combining 3D and 2D group convolution [17]. This approach alleviates the dense 3D convolution cost while maintaining the benefits of cross-modality features. We also employ a channel shuffle operation after each cross-modality convolution layer to enhance information flow within groups, inspired by [29].Our dataset, collected from a Taiwan hospital, includes 88 stage-one and 50 stage-two patients, each with three MR image types: T1-weighted, T2-weighted, and Out-of-phase images. We thoroughly evaluate our framework in small dataset settings and demonstrate its effectiveness in liver cancer classification. Our major contributions can be summarized as follows:

- We present an effective and efficient training framework to classify the pathological stage of liver cancer from multi-modality MR images, which can assist doctors in disease diagnosis.
- We propose multi-modality data augmentation, an effective approach to adapt contrastive learning into multi-modality learning.

- We introduce a lightweight fusion scheme for multi-modality features by combining cross-modality group convolution and the channel shuffle operation.
- The proposed framework can leverage multi-modality data to further enhance the robustness of the model both in the pre-training and fine-tuning stage.

2 RELATED WORK

Contrastive Learning

Contrastive learning has shown promising results in computer vision [3, 4, 7], where positive pairs of latent features should be similar and negative pairs dissimilar [2, 27]. BYOL and SimSiam are effective Siamese architectures in contrastive learning [3, 4, 7], with BYOL also resilient to batch size and data augmentation changes. However, in single-modality scenarios, positive pair construction often relies on random augmentations of the same view, limiting the model's ability to incorporate mutual information between modalities.

In video representation learning, positive pairs are selected from adjacent frames [24], and MOCO is adapted for chest X-ray detection [21]. Various strategies for positive pair selection have been explored [1, 24, 26]. We address the multi-modality scenario and improve pre-training methods by using MDA to adapt contrastive learning for learning multi-modality representations. While inspired by previous works [1, 22], we also further leverage contrastive representations during fine-tuning.

Multi-modality Learning

Several works apply the generative adversarial network (GAN) [6] to multi-modality learning. For example, Jiang et al. [12] uses GANs to generate synthetic data as augmented data for joint training. Li et al. [18] propose an image align module to fill the appearance gap between modalities and make use of cross-modality information to assist segmentation tasks on target-modality via knowledge distillation. This work explores the prior knowledge of assistant-modality to enhance the performance on target-modality, while we aim to learn multi-modality representations jointly and further exploit multi-modality features to boost the overall effectiveness. Tseng et al. [25] adapt a dense 3D convolution layer with a convolutional LSTM to model multi-modality features and sequence slices. To incorporate local and global features, Kamnitsas et al. [14] propose multi-scale 3D CNN with a dual pathway structure. Fidon et al. [5] present a nested structure, which is scalable to the number of input modalities to leverage multi-modality features. Guo et al. [8] analyze different fusion schemes for multi-modality images and demonstrates that feature fusion at the feature level is superior to the decision-making level and the classifier level. The studies mentioned above focus on the design of the feature extractor or different schemes of the feature fusion structure. However, we propose to use the self-supervised model as a weight-sharing feature extractor. Finally, we adopt an efficient and effective feature fusion strategy for the contrastive multi-modality features.

3 METHODOLOGY

In line with practices in natural language processing and computer vision, we adopt a self-supervised model to leverage structural information between MR images. Siamese networks facilitate effective visual representation learning during pre-training, as seen in successful methods like [4, 7].

During fine-tuning, the pre-trained model serves as a weight-sharing multi-modality encoder, leading to significant reductions in parameters and computation costs. We propose further utilizing multi-modality features with a lightweight cross-modality convolution in the downstream task (see Fig 2).

Algorithm 1 Multi-modality Contrastive Learning

Input: Patients sets $\mathcal{D}$ and batch size N
Augmentation and preprocessing functions $\mathcal{T}$
Networks of f, g, h
Stop gradient operation sg

while not converge **do**
  $\mathcal{B} \leftarrow \{p_i \sim \mathcal{D}\}_{i=1}^{N}$
  # sample minibatch of patients
  for $p_i \in \mathcal{B}$ **do**
    $t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$
    # draw transformations
    draw positive pairs $x_1 \sim p_i$ and $x_2 \sim p_i$
    $x_1' = t(x_1)$ and $x_2' = t'(x_2)$
    $q_1 \leftarrow g(f(x_1'))$ and $q_2 \leftarrow g(f(x_2'))$
    # projections
    $z_1 \leftarrow h(q_1)$ and $z_2 \leftarrow h(q_2)$
    # predictions
    define $\mathcal{D}(z, q)$ **as** $\mathcal{D}(z, q) = 2 - 2 \cdot \frac{z \cdot q}{\|z\|_2 \cdot \|q\|_2}$
    # mse loss
    $\ell_i \leftarrow \mathcal{D}(z_1, sg(q_2)) + \mathcal{D}(z_2, sg(q_1))$
  end for
  $\mathcal{L} = \frac{1}{N} \sum_{k=1}^{N} \ell_i$
  update networks f, g and h to minimize $\mathcal{L}$
end while
return networks f, g

3.1 Multi-modality Contrastive Learning

Inspired by SimSiam architecture, our pre-train flow is shown in Figure 1. The proposed multi-modality data augmentation can apply to various contrastive learning architectures, such as SimCLR and BYOL. Multi-modality data augmentation increases the difficulty of the pretext task and encourages the model to learn a cross-modality representation. We randomly select two MR images x_1 and x_2 from a patient as positive pairs. Since different modalities have different image statistics (e.g., mean or standard deviation), we apply standard normalization for each modality. We then take two random augmentations for two samples. The augmented images x_1' and x_2' will be encoded by the same encoder f and projector g, which are $q_1 \triangleq g(f(x_1'))$ and $q_2 \triangleq g(f(x_2'))$ respectively. The predictor h will project q_1 to map to its positive pair q_2, in which $z_1 \triangleq h(q_1)$. Finally, we minimize the mean square error between two normalized vectors $\overline{z_1}$ and $\overline{q_2}$ [7], where $\overline{z_1} \triangleq z_1/\|z_1\|_2^2$ and $\overline{q_2} \triangleq q_2/\|q_2\|_2^2$,

$$\mathcal{D}(z_1, q_2) = \|\overline{z_1} - \overline{q_2}\|_2^2 = 2 - 2 \cdot \frac{z_1}{\|z_1\|_2} \cdot \frac{q_2}{\|q_2\|_2} \tag{1}$$

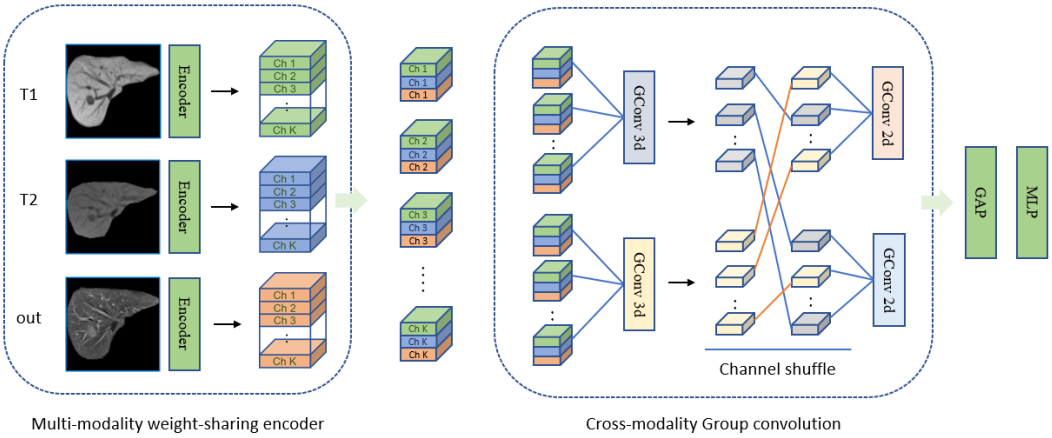

Figure 3: An illustration of the proposed overall architecture. The self-supervised pre-trained model serves as a multi-modality encoder. In the downstream task, we use the CGC layer by combining Conv1 and Conv2 block to incorporate multi-modality features.

Symmetrized the loss [7] by swapping the input images to corresponding networks, we define symmetrized loss as:

$$\mathcal{L} = \mathcal{D}(z_1, q_2) + \mathcal{D}(z_2, q_1) \tag{2}$$

To prevent collapsing to trivial solutions (e.g., the network produces the same output for all images), the stop-gradient operation [4] is added to the right side of the pre-trained flow in Figure 1:

$$\mathcal{L} = \mathcal{D}(z_1, stopgrad(q_2)) + \mathcal{D}(z_2, stopgrad(q_1)) \tag{3}$$

The stop gradient operation will treat the corresponding valueas a constant. Thus, no gradient will backpropagate along with the corresponding value. This operation is a central component of networks to prevent collapse solutions. At the end of the training, the networks f and g are returned for downstream use.

3.2 Cross-modality Group Convolution

In this section, we introduce how to aggregate multi-modality representative features. Inspired by the work [25], we propose a lightweight cross-modality group convolution (CGC) to make efficient use of multi-modality features. The overall architecture is shown in Figure 3, the strategy of feature fusion focuses on information flow between or within channels of the multi-modality feature map. Specifically, we have three kinds of images for each patient, and each image will be encoded in a feature map of size $C \times H \times W$. Following the similar idea from the work [25], we stack together the same channels of different modalities features, in which the feature size is $C \times 3 \times H \times W$. We then perform $3 \times 1 \times 1$ group convolution followed by three layers of 1×1 group convolution (which is called pointwise group convolution [29]). Group convolution was first introduced in AlexNet [17] for distributed training. Here, the group convolution is used to reduce computation cost by ensuring each convolution of a group operates only in the corresponding channels. However, this design decreases the number of information flows across groups since the input of a certain group is only connected to the output of a certain group. To enhance communication between channel groups, the channel shuffle operation [29] is applied to each output of the group convolution layer. This is implemented by reshaping the channel dimension to $g \times n$, in which g is the number of groups and n is the number of group channels. The output channel dimension is then transposed and flattened, and is finally input into the next convolution layer. This operation is differentiable and can be embedded in networks. Group convolution layer with channel shuffle allows more feature map channels, and is crucial for efficiently aggregatating the multi-modality feature map.

4 EXPERIMENTS

4.1 Datasets and Experimental settings

Our dataset comprises three MRI modalities: Out-of-phase, T2-weight, and T1-weight imaging, each representing a transverse section of the thoracic cavity. Collected from a hospital in Taiwan, it consists of 414 MR images from 138 patients, with 88 at stage one and 50 at stage two. All MR images are liver region segmented. Inspired by the effectiveness of ImageNet pre-trained models in the medical domain [1, 15], we use ImageNet pre-trained weights in all our experiments. Pre-training and fine-tuning stages share the same preprocessing methods but differ in data augmentation. The choice of data augmentations significantly affects contrastive learning effectiveness [2, 7]. To improve the quality of representations, the pretext task should be challenging [2]. We employ SimCLR with random cropping and color jitter to prevent shortcut exploitation in pretext tasks, while SimSiam and BYOL are receptive to combinations of data augmentations. Although random cropping may alter image labels in the medical domain, our focus is on capturing the

Table 1: Linear evaluation results on various pre-training methods. Multi-modality data augmentation (MDA denoted as +) is applied in SimCLR, BYOL, and SimSiam. The values are averaged from different random seeds and reported in %.

	Out				T2				T1			
Methods	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
SimCLR	64.22	52.09	46.80	48.41	62.72	50.14	44.67	46.46	62.12	49.52	43.20	45.44
SimCLR+	**65.95**	**54.37**	**47.14**	**49.10**	**64.20**	**50.18**	**45.20**	**46.88**	**64.02**	**52.53**	**48.00**	**49.03**
BYOL	66.44	55.52	48.29	50.52	66.76	54.34	49.33	51.27	65.48	52.54	49.60	50.60
BYOL+	**67.73**	**55.91**	**57.14**	**55.63**	**68.28**	**58.29**	**53.71**	**54.86**	**66.88**	**58.86**	**51.33**	**51.16**
SimSiam	66.87	54.95	53.20	53.37	65.17	53.85	47.20	48.26	63.06	51.21	45.60	47.56
SimSiam+	**68.88**	**57.95**	**54.36**	**55.34**	**67.47**	**55.23**	**52.91**	**53.51**	**66.57**	**55.31**	**52.73**	**53.09**

relationship between modalities during pre-training, evaluating on the liver cancer classification task.

For evaluation, we use five-fold cross-validation with five different random states to assess stability with varied data combinations.

4.1.1 Implementation details for pretraining. We use EfficientNet-B2 [23] without multi-layer perceptron (MLP) as our base encoder. The weight of the encoder is initialized by ImageNet pre-trained. The projector consists of a global average pooling layer with dimension 1,408 and a 2-layer of MLP with hidden size 1,792 and output size 256. Batch normalization (BN) [11] and sigmoid Linear Unit (Swish) [20] have been applied to hidden layers. The predictor is a 2-layer multi-layer perceptron (MLP) with the same architecture as the projector, except for the global average pooling layer. Using the idea in [7], the output of both predictor and projector are not batch normalized. Data augmentation is applied in this stage, and technologies include color jitter, random crop, horizontal and vertical flip. All images are resized to 200×200 for training. We apply distinct mean and standard deviation for normalizing each modality. The mean value is set to 0.2962, 0.2055, and 0.2000 for T1 HB, T2, and out modality respectively. The standard deviation is set to 0.2527, 0.1727, and 0.1644 for T1 HB, T2, and out modality respectively. For smaller images, zero-padding is used instead of interpolation [9]. We use Adam optimizer for pre-training over 12 epochs. The initial learning rate is set to $3 \cdot 10^{-4}$ with batch size 32.

4.1.2 Implementation details for CGC. Swish and BN have been applied to all group convolutions layers before the channel shuffle operation. Using the idea from previous works [10, 29], the shortcut connection is inserted between every two consecutive convolution blocks. The number of groups for all convolutions layers is set to 8. Global average pooling and an MLP are performed after multi-modality feature fusion. The input dimension of MLP is 1,408 with a dropout rate of 0.3. We use Adam optimizer to train over 10 epochs. The initial learning rate is $1 \cdot 10^{-4}$ with a linear warm-up period of 50 steps, and the batch size is 16. In this stage, our data augmentation includes color jitter, random rotation, and random flipping. The images are resized or padded to 200×200. Normalization is performed after the data augmentation. We emphasize that the multi-modality data augmentation can create a more comprehensive view of images.

Table 2: Semi-supervised training with different weight initializations using SimSiam architecture for single-modality (SCL) and multi-modality contrastive learning (MCL). Averaged results from 5 different random seeds. (±denotes standard deviation)

	Out		T2		T1	
Methods	Acc	F1	Acc	F1	Acc	F1
ImageNet	67.51 ±1.16	54.11 ±2.46	65.27 ±1.43	55.42 ±2.05	66.53 ±1.69	52.24 ±1.16
SCL	69.07 ±1.22	58.13 ±2.81	67.43 ±1.71	55.51 ±1.41	67.55 ±1.24	55.04 ±2.14
MCL	**72.02 ±1.98**	**62.81 ±1.93**	**70.26 ±0.68**	**58.83 ±2.66**	**69.86 ±0.57**	**58.14 ±2.27**

4.2 Linear evaluation

Following the widely used evaluation procedure [2, 7, 16, 28], we evaluate the generalization capability of the self-supervised pre-trained model. Specifically, we train a linear classifier on the top of the frozen base network to classify liver cancer. The network consists of the base encoder f and projector g, which are jointly pre-trained with three types of MR images using MDA. Without MDA to incorporate MR images, the network will be trained separately on each type of MR images, relying on data augmentation to form positive pairs. To assess the impact of multi-modality information incorporation, we adopt three methods to measure the impact of MDA, including SimCLR, BYOL, and SimSiam. The cross-validation results, which are averaged from different random seeds and shown in Table 1. The result of SimCLR reports that its architecture critically depends on a large batch size [2], which has deteriorative performance according to our experiment. With MDA applied to each architecture, the linear evaluation results show that representation quality will be improved when including all kinds of MR images. We conclude that the pre-trained encoder with MDA benefits from the shared information across multi-modality MR images.

4.3 Semi-supervised learning

To further evaluate the learned representations, we access the performance when fine-tuning the self-supervised model on liver cancer classification. Here, the architecture of networks is the base encoder f and projector g pre-trained from SimSiam networks and an extra MLP. The base encoder is pre-trained from MR images of all patients with MCA and pre-trained from a single modality of

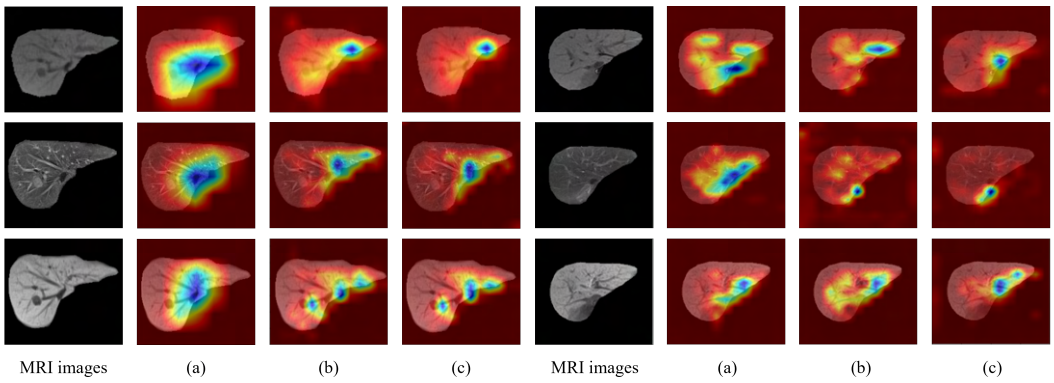

Figure 4: Class activation heat map on three types of MR images under different trained models. We compare the model between the (a) Pre-trained MCL, (b) Fine-tuned MCL, and (c) Fine-tuned MCL with CGC.

all patients with SCA. In the fine-tuned stage, we split the dataset according to patient ID. We compare the representations quality trained from single-modality contrastive learning (SCL) to multi-modality contrastive learning (MCL). The fine-tuned results on three types of MR images are shown in Table 2. The accuracy of fine-tuned SCL is greater on three modalities (1.56% for Out-of-phase images, 2.16% for T2-weighted images, and 1.02% for T1-weighted images) than supervised pre-training on ImageNet. This is consistent with previous studies in the medical domain [1, 15, 26]. These results are attributed to self-supervised pre-training in the task-specific unlabeled data that can bridge the domain discrepancy. Besides, the MCL model improves more on three types of MR images, improving the accuracy of SCL with 2.95% for Out-of-phase images, 2.83% for T2-weighted images, and 2.31% for T1-weighted images. We conclude that the fine-tuned MCL model can benefit from cross-modality information and has better generalization ability. Finally, we evaluate the performance obtained from CGC trained on top of base encoder f. Experimental results are reported in Table 3. In this stage, the ImageNet pre-training followed by multi-modality contrastive learning (MCL) is our weight initialization method. Compared to the previous works [1, 26], which focus on the pre-training strategy, we further exploit the representative features learned from different imaging modalities using the proposed CGC. The results show the significant improvement of CGC by leveraging cross-modality information. The CGC improves the accuracy on Out-of-phase, T2-weight, and T1-weight images with 4.78%, 6.54%, and 6.94%, respectively. For both the linear evaluation and fine-tuning stage, MCL performs better than SCL, since MCL can learn meaningful multi-modality representations. Moreover, we propose that the incorporation of different multi-modality images can improve the quality of representations. Our experiment verifies that the proposed CGC can further capture cross-modality information and boost the robustness of the classification.

Additionally, to demonstrate the effectiveness and efficiency of the proposed CGC compared to the cross-modality convolution (CMC)[25], we conduct an experiment to compare the classification performance and complexity between CMC and CGC, and the averaged results from different random states are shown in Table 4. Under the same pre-trained method and the base encoder, the

Table 3: Comparison of fine-tuned MCL performance with and without cross-modality group convolution (CGC) on each type of MR image. Averaged results from 5 different random seeds. (±denotes standard deviation)

Methods	Acc	Pre	Rec	F1
MCL + Out	72.02 ±1.72	62.71 ±3.28	67.18 ±3.14	62.81 ±1.93
MCL + T2	70.26 ±0.68	58.85 ±4.58	63.10 ±3.57	58.83 ±2.66
MCL + T1	69.86 ±0.57	60.06 ±1.80	60.89 ±5.56	58.14 ±5.28
MCL + CGC + ALL	**76.80 ±1.40**	**71.11 ±5.16**	**70.59 ±2.63**	**67.82 ±1.59**

Table 4: Performance and Complexity comparison

Methods	Acc	F1	Params	Flops
CMC	74.31	66.04	13.69M	1.29B
CGC	76.80	67.82	9.19M	1.07B

accuracy degraded by 2.49% when replacing the CGC to CMC. We demonstrate that both the classification performance and complexity of our method can outperform previous methods in liver cancer classification. Within the desired computation budget, CGC allows more feature map channels which can aggregate multi-modality information and is invaluable for small datasets.

4.4 Visualization

To explain the result, we visualize the class activation maps on liver MR images using LayerCAM [13] shown in Figure 4. LayerCAM can generate more fine-grained object localization information from the class activation maps, using the gradients to highlight the important region in the feature map. The level of importance from high to low is in the following order: blue, green, red. We choose the 11th convolution block of the trained base encoder as our target layer and visualize pathological stage two for the target class. We note that the MCL pre-trained model still has lots of noise. After fine-tuning, the model tends to highlight the specific position. The final results of MCL with the CGC layer demonstrate the effectiveness of our idea. The class activation map shows less noise since our

method can aggregate the cross-modality information. Moreover, our model concentrates on the underlying pathological locations. We conclude that the CGC can incorporate cross-modality features to enhance the robustness of the model.

5 CONCLUSIONS

In this work, we introduce an effective and efficient training framework based on contrastive learning for liver cancer classification. We present multi-modality data augmentation as a simple approach to adapt contrastive learning into multi-modality learning. More importantly, we exploit the representative multi-modality features by using a lightweight cross-modality group convolution to integrate multi-modality information. For practical purposes, we also visualize the model to help the doctor comprehend the model prediction. Experimental results on datasets collected from the real-world demonstrate the effectiveness of our methods.

ACKNOWLEDGMENTS

This work was partially supported by National Science and Technology Council (NSTC) under Grants 111-2636-E-006 -026 -, 112-2221-E-006 -100 - and 112-2221-E-006 -150 -MY3. The authors are grateful to Chi Mei Medical Center for providing the medical dataset.

REFERENCES

[1] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. 2021. Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224* (2021).

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[3] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029* (2020).

[4] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15750–15758.

[5] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sebastien Ourselin, and Tom Vercauteren. 2017. Scalable multimodal convolutional networks for brain tumour segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 285–293.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733* (2020).

[8] Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. 2018. Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 903–907.

[9] Mahdi Hashemi. 2019. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. *Journal of Big Data* 6, 1 (2019), 1–13.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

[12] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S Mageras, Joseph O Deasy, and Harini Veeraraghavan. 2018. Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 777–785.

[13] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. 2021. LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Transactions on Image Processing* 30 (2021), 5875–5888.

[14] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis* 36 (2017), 61–78.

[15] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, and Pranav Rajpurkar. 2021. CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*. 116–124.

[16] Nikos Komodakis and Spyros Gidaris. 2018. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[18] Kang Li, Lequan Yu, Shujun Wang, and Pheng-Ann Heng. 2020. Towards cross-modality medical image segmentation with online mutual knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 775–783.

[19] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[20] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941* (2017).

[21] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. 2020. MoCo-CXR: MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models. *arXiv preprint arXiv:2010.05352* (2020).

[22] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. 2021. Multimodal self-supervised learning for medical image analysis. In *International Conference on Information Processing in Medical Imaging*. Springer, 661–673.

[23] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[24] Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. 2020. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13806–13815.

[25] Kuan-Lun Tseng, Yen-Liang Lin, Winston Hsu, and Chung-Yang Huang. 2017. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6393–6400.

[26] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. 2021. MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. *arXiv preprint arXiv:2102.10663* (2021).

[27] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*. PMLR, 9929–9939.

[28] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *European conference on computer vision*. Springer, 649–666.

[29] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6848–6856.