



HAL
open science

Reinforcement Learning Driven Intra-modal and Inter-modal Representation Learning for 3D Medical Image Classification

Zhonghang Zhu, Liansheng Wang, Baptiste Magnier, Lei Zhu, Defu Zhang,
Lequan Yu

► **To cite this version:**

Zhonghang Zhu, Liansheng Wang, Baptiste Magnier, Lei Zhu, Defu Zhang, et al.. Reinforcement Learning Driven Intra-modal and Inter-modal Representation Learning for 3D Medical Image Classification. MICCAI 2022 - The 25th International Conference on Medical Image Computing and Computer Assisted Intervention, Sep 2022, Singapour, Singapore. pp.604-613, 10.1007/978-3-031-16437-8_58. hal-03790158

HAL Id: hal-03790158

<https://imt-mines-ales.hal.science/hal-03790158v1>

Submitted on 4 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Reinforcement Learning Driven Intra-modal and Inter-modal Representation Learning for 3D Medical Image Classification

Zhonghang Zhu¹, Liansheng Wang¹(✉), Baptiste Magnier², Lei Zhu^{3,4}, Defu Zhang¹, and Lequan Yu⁵

¹ Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China, zzhonghang@stu.xmu.edu.cn, {lswang,dfzhang}@xmu.edu.cn

² Euromov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France, baptiste.magnier@mines-ales.fr

³ The Hong Kong University of Science and Technology (Guangzhou), China,

⁴ The Hong Kong University of Science and Technology, Hong Kong SAR, China, leizhu@ust.hk

⁵ Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China, lqyu@hku.hk

Abstract. Multi-modality 3D medical images play an important role in the clinical practice. Due to the effectiveness of exploring the complementary information among different modalities, multi-modality learning has attracted increased attention recently, which can be realized by Deep Learning (DL) models. However, it remains a challenging task for two reasons. First, the prediction confidence of multi-modality learning network cannot be guaranteed when the model is trained with weakly-supervised volume-level labels. Second, it is difficult to effectively exploit the complementary information across modalities and also preserve the modality-specific properties when fusion. In this paper, we present a novel Reinforcement Learning (RL) driven approach to comprehensively address these challenges, where two Recurrent Neural Networks (RNN) based agents are utilized to choose reliable and informative features within modality (intra-learning) and explore complementary representations across modalities (inter-learning) with the guidance of dynamic weights. These agents are trained via Proximal Policy Optimization (PPO) with the confidence increment of the prediction as the reward. We take the 3D image classification as an example and conduct experiments on a multi-modality brain tumor MRI data. Our approach outperforms other methods when employing the proposed RL-based multi-modality representation learning.

Keywords: Multi-modality Learning · 3D Medical Images · Reinforcement Learning · Classification.

1 Introduction

Multi-modality images, *e.g.*, different MRI, are widely used in medical applications [2,18]. Integrating the strengths of multiple modalities by exploring their

rich information and discovering the underlying correlations among them is an effective manner to improve the diagnosis and prognosis tasks. In other aspects, many medical images involve 3D format, therefore, multi-modality 3D image classification is important in medical image computing field. However, it is challenging to develop such multi-modality learning algorithms for several reasons. On the one hand, for learning within a single modality (*i.e.*, intra-modal learning), since obtaining slice-level labels of a 3D volume via manual labeling is tedious and time-consuming [6]; it is difficult to identify features containing modality-specific information without precise instructions. On the other hand, for learning among different modalities (*i.e.*, inter-modal learning), since the underlying correlations among them are unclear, exploiting complement yet discriminative representations from each modality is also non-trivial.

Recently, an increasing number of studies have been investigated for multi-modality learning. As an example, to obtain complementary features of different modalities, Canonical Correlation Analysis (CCA) [3] projects the features of each modality to a new robust space. Multiple Kernel Learning (MKL) [7] utilizes a set of predefined kernels from multi-view data to integrate these modalities using the optimized weights. In addition, there are several works that applied DL networks for multi-modal learning [20,11,15,13,9]. These methods can be roughly categorised into two branches: 2D-based methods [19,10] and 3D-based methods [5,8,14]. For the first line of methods, 3D volumes are firstly projected into 2D images and then are integrated for the final prediction. However, these methods are insufficient in capturing the complicated spatial characteristics of 3D volumes. In contrast, 3D-based methods can work well in capturing spatial relations between different volumes for learning more complementary multi-modality representations. However, there still exist several limitations for 3D-based methods. First, the particular use of the 3D fusion models can only be supervised by the volume-level labels with limited information, which leads to high uncertainty prediction [1]. While the risk-sensitive tasks, like medical diagnosis, require high prediction confidence for the purposes of avoiding critical mistakes. Second, to explore complementary representation in multi-modality learning, the weighted fusion method with fixed weight is widely applied in these 3D fusion researches. However, it is unreasonable to merely assign specific weights to different modalities, besides, the weights should be dynamically allocated by data-driven rather than artificial.

In this paper, a novel Reinforcement Learning (RL) driven approach for effective multi-modality 3D medical image analysis is presented, where two RL-based agents are learned for dynamical intra-modal and inter-modal features enhancement to learn latent modality representation and underlying correlations among different modalities. Specifically, to enable such intra-modal and inter-modal feature enhancement, we explore two key techniques based on the characteristics of multi-modality medical images. (1) We propose an iterative hybrid-enhancement network to integrate intra-features and inter-features, where the enhanced intra-features in each iteration are regarded as the state of two designed agents which generate strategies for intra-enhance and inter-enhance in the next training iter-

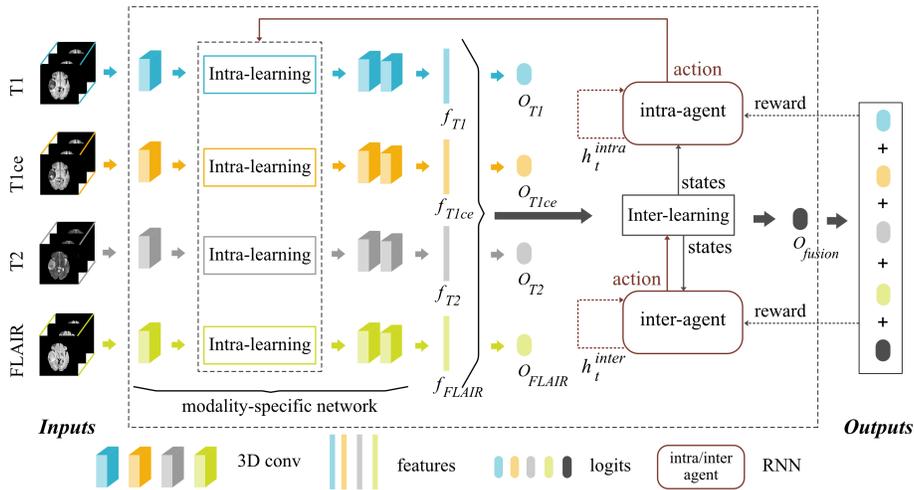


Fig. 1. Illustrations of the proposed method: given a group of 3D volumes in different modalities *i.e.*, T1, T1ce, T2, FLAIR, the model iteratively processes a sequence of multi-modality images with intra-learning and inter-learning modules. The proposal agents (intra-agent and inter-agent) are supervised by the reward function designed as confidence increment of the outputs. Hidden states h_t^{intra} and h_t^{inter} of both agents help them exploit information from previous inputs and produce promising actions for intra-modality learning and inter-modality learning, respectively.

ation. (2) We take the prediction confidence increment as the supervision of the agents, which encourages the agents to customize the enhanced strategies that can promote the prediction confidence. Finally, the whole framework is trained in an end-to-end manner for hybrid multi-modality learning.

2 Method

The detailed architecture of our proposed framework is shown in Fig. 1. Given an input volume, in order to capture intra-modality representations, a 3D convolution layer in the modality-specific network is first employed to generate shallow modality features which are enhanced by the RL-based intra-learning module to focus on the salient part. Then, modality-specific features are passed to another RL-based inter-modal learning module to conduct multi-modality learning for prediction promotion. Specifically, the output of the agent is taken as the state of the intra-agent and inter-agent to determine actions (enhancement weights) for intra-/inter- modality learning in the next iteration. The increments of the softmax prediction are used as reward function to train the agents for facilitating the agents to propose actions that enable the network to produce correct predictions in high confidence. The detailed design of RL-based intra-/inter- modality learning process is introduced in the following subsections.

2.1 RL-Based Intra-modality Learning

In the proposed method, the RL based intra-modality learning is an alternating decision making process. Specifically, at the beginning of a training iteration, 3D volumes sized in $B \times 1 \times d \times w \times h$ are sent to individual 3D convolution layers to get the modality-specific primary representations sized in $B \times c \times d \times w \times h$, respectively. The B , c denotes the batch size and channel number, while d , w , h indicates the depth, width and height of the input 3D volumes, respectively. Intra-modality learning is then implemented by multiplying the intra-modality enhancement weights $W_{s,t} \in \mathbb{R}^{B \times c}$ with modality-specific primary representations, where t denotes the number of iterations.

We extract the feature from the shallow layer of the modality-specific network to conduct intra-modality learning with the assumption that these features keep the modality-specific spatial relationship. Moreover, different modalities share the same intra-modality enhancement weights determined by the intra-agent whose initial state is designed as the sum of different high-level modality-specific representations, *i.e.*, $f_{T1} \in \mathbb{R}^{B \times 8c}$, $f_{T1ce} \in \mathbb{R}^{B \times 8c}$, $f_{T2} \in \mathbb{R}^{B \times 8c}$, $f_{FLAIR} \in \mathbb{R}^{B \times 8c}$, aiming that the proposed intra-enhancement action can facilitate sharing of intra-modality spatial relationships among different modalities. In our experiments, intra-modality enhancement weights $W_{s,t}$ in the first iteration are initialized as 1, and updated by the intra-agent in the following iterations.

2.2 RL-Based Inter-modality Learning

In inter-modality learning, different high-level modality-specific representations f_{T1} , f_{T1ce} , f_{T2} , f_{FLAIR} are fused to generate a $B \times 32c$ feature forwarded to two independent RNN termed intra-agent and inter-agent. Inter-modality enhancement weights $W_{f,tk} \in \mathbb{R}^{B \times 1}$, $k \in (0, 1, 2, 3)$ are involved in the inter-modality learning which can be formulated as:

$$f_{fusion} = Concat(f_{T1} \times W_{f,t0}, f_{T1ce} \times W_{f,t1}, f_{T2} \times W_{f,t2}, f_{FLAIR} \times W_{f,t3}), \quad (1)$$

where the *Concat* represents the concatenation operation, and $W_{f,t0}$, $W_{f,t1}$, $W_{f,t2}$, $W_{f,t3}$ are different weights for different modalities. Same as intra-modality enhancement weights $W_{s,t}$, the inter-modality enhancement weights $W_{f,tk}$ are also set as 1 when $t = 0$, and subsequently updated by the inter-agent.

With the fusion feature f_{fusion} , a Fully Connected (FC) layer is adopted to generate a primary fusion representation. The primary fusion representation includes comprehensive features of all modalities, which is suitable to be set as the initial state of the inter-agent. While if we split the fusion feature f_{fusion} into individual modality-specific features and add them together, the voting information of each modality-specific feature for the current predicted state can be synthesized to the greatest extent, without harming characteristics specific of each modality, which can be regarded as states of the intra-agent. with the initial states, the intra-agent and inter-agent will be triggered to propose the new $W_{s,t}$ and $W_{f,tk}$, $k \in (0, 1, 2, 3)$ for next iteration procedure.

Note that several iterations are conducted in a training epoch. For each iteration, the hidden state h_t^{intra} and h_t^{inter} which aggregates the information of past states are maintained within intra-agent and inter-agent, respectively. In addition, modality-specific logit outputs O_{T1} , O_{T1ce} , O_{T2} and O_{FLAIR} , *i.e.*, the output of each modality-specific network followed a FC layer and a soft-max layer, with modality fusion logit O_{fusion} are saved in a memory bank until the end of the epoch. These sequential logit outputs will be used for calculating reward of the agents and the training loss.

2.3 Reward Function and Training Procedure

Operating under the intuition that the normalized prediction probability (*i.e.*, normalized from 0 to 1) reflects model confidence of the prediction. In this study, we define the reward as increments of the soft-max prediction probability on the ground truth labels for the optimization of two agents. Specifically, the reward function to train the agent is designed as $r_t = p_t - p_{t-1}$, where p_t is the soft-max prediction probability with the ground truth label at the t^{th} iteration process, which is derived from the t^{th} logit outputs. Then both the inter-agent and intra-agent are trained simultaneously by maximizing discounted reward $r_{all} = \sum_{t=1}^3 \gamma^{t-1} \cdot r_t$ where the γ is set to 0.1.

The network of our RL-based multi-modality learning model includes two components: modality-specific network consisted of four cascaded 3D encoders which are used for representation extraction of each single-modality, and a RL module includes two agents for intra-/inter- modality enhancement weights proposal. Each 3D CNN extractor contains six 3D blocks followed the avgpool layer and FC layer. Besides, the inter-agent and intra-agent share a similar structure including a RNN followed a FC layer, but the FC layer has different nodes number which is set as c for intra-agent and 4 for inter-agent, respectively.

To optimize the whole framework, we collect modality-specific logits and fusion logits of each iteration to calculate the training loss. Subsequently, the total loss of our approach can be computed as:

$$L = \sum_{t=0}^3 L_{O_{T1},t} + L_{O_{T1ce},t} + L_{O_{T2},t} + L_{O_{FLAIR},t} + \lambda \cdot L_{O_{fusion},t}, \quad (2)$$

where $L_{O.,t}$ denotes loss calculated by cross entropy loss at the t^{th} iteration and $L_{O_{T1},.}$, $L_{O_{T1ce},.}$, $L_{O_{T2},.}$, $L_{O_{FLAIR},.}$ and $L_{O_{fusion},.}$ represent the loss calculated between outputs of each iteration and the labels. The λ is set to 4 as a trade-off parameters to balance the influence of modality-specific loss and fusion loss. In the test phase, we only keep the sum of the outputs of the last iteration as the final outputs which can be denoted as $Outputs = O_{T1,3} + O_{T1ce,3} + O_{T2,3} + O_{FLAIR,3} + O_{fusion,3}$.

Table 1. Quantitative results (mean±standard deviation) of different methods on BraTS18.

| Method | Acc | Precision | Recall | F1 score |
|-------------------------|--------------------|--------------------|--------------------|--------------------|
| 2D CNN | 0.613±0.178 | 0.617±0.259 | 0.581±0.190 | 0.555±0.224 |
| TPCNN [17] | 0.636±0.000 | – | – | – |
| M ² Net [19] | 0.664±0.061 | 0.574±0.141 | 0.613±0.075 | 0.589±0.102 |
| 3D CBAM [16] | 0.650±0.087 | 0.603±0.082 | 0.516±0.077 | 0.518±0.084 |
| 3D MFB [18] | 0.575±0.018 | 0.586±0.022 | 0.540±0.017 | 0.521±0.037 |
| Ours | 0.692±0.189 | 0.675±0.251 | 0.648±0.205 | 0.622±0.242 |

3 Experiment

3.1 Experiment Setup

Datasets. Experiments are carried out on the BraTS 2018: a multi-modality MRI dataset in which each patient includes T1, T1ce, T2 and FLAIR volumes. In this study, we also focus on the overall survival prediction task defined in [19], finally, we have 165 subjects with survival information for this dataset. Our prediction task is constructed following [19], in which patients are divided into three classes: (1) low survival risk, (2) middle survival risk, (3) high survival risk.

Implementation Details. For experiments of BraTS 2018, we first locate the tumor region according to the tumor mask and extract an image volume that is centered on the tumor region. Then we resize the extracted image volume of each subject to a predefined size (*i.e.*, $d = 64$, $w = 64$, $h = 64$). Three individual Adam optimizers are taken to train the feature extraction backbone and two agents, respectively. The learning rate of the feature extraction backbone is set as $1e-4$ while the learning rate of agents is set as $1e-2$ and the weight decays are set to $1e-5$. In every training epoch, the agent will iterate three times ($t=3$). The batch size is set to 10 and we adopt a 10-fold cross-validation and report the average performance of 10 folds. For each fold, we further divide the data (the other 9 folds) into training set (80%) and validation set (20%) and take the best model on validation part for evaluation.

Evaluation Metric. We evaluate our method with Accuracy (Acc), Precision, Recall and F1 score. The precision and recall are calculated with one-class-versus-all-other-classes and then calculate F1 score ($F1 = \frac{2*Precision*Recall}{Precision+Recall}$).

3.2 Experimental Results

Comparisons with the State of the Art. The proposed method is also compared with other fusion methods, results are reported in Table 1. For comparisons, we choose the following methods: 1) 2D CNN fusion: we project the input 3D volume onto 2D images along the vertical axis by averaging the sum of all slices for each modality. Then, we use the ResNet34 [4] (replace the input

Table 2. Quantitative results (mean±standard deviation) of ablation studies on BraTS18.

| Method | Acc | Precision | Recall | F1 score |
|----------------|--------------------|--------------------|--------------------|--------------------|
| Baseline | 0.631±0.081 | 0.546±0.087 | 0.546 ±0.043 | 0.574±0.088 |
| Ours w/o Inter | 0.681± 0.194 | 0.666±0.259 | 0.660±0.210 | 0.627±0.234 |
| Ours (1 modal) | 0.600±0.071 | 0.462±0.095 | 0.429±0.063 | 0.364±0.033 |
| Ours (2 modal) | 0.614 ±0.065 | 0.558±0.026 | 0.496±0.015 | 0.431±0.035 |
| Ours (3 modal) | 0.650±0.079 | 0.620±0.062 | 0.575±0.019 | 0.528±0.039 |
| Ours | 0.692±0.189 | 0.675±0.251 | 0.648±0.205 | 0.622±0.242 |

channel from three to one) as the modality-specific feature extractor, and then fuse the outputs using the concatenation operation. 2) TPCNN [17]: this method uses a CNN model to extract features from multi-modal data and then employs XGBoost to build the regression model. 3) M²Net [19]: a multi-modal shared network to fuse modality-specific features using a bilinear pooling model, exploiting their correlations to provide complementary information. 4) 3D CBAM [16]: using the CBAM [16] module to adaptively generate enhancement weights for intra-/inter- modalities. 5) 3D MFB [18]: a hybrid-fusion network with Mixed Fusion Block (MFB) to adaptively weight different fusion strategies. We retain the encoder of the modality-specific network with the MFB block in [18] to match our classification task. T1 and FLAIR modalities are taken as inputs.

Note that we have the same setting as the TPCNN and M²Net, and take the results reported in [19]. As shown in Table 1, the 3D CBAM [16] which fuses features under channel attention mechanism in the training process, achieves worse performance than the proposed method. In addition, the 3D CBAM shares the same baseline network with ours, indicating that the accuracy boost is due to the RL-agent module not the increasing of backbone size. Moreover, we find that the agent tends to pay more attention to the FLAIR modality from the learned weights for inter-modality fusion. It proves that the proposed RL-based hierarchic multi-modality learning method can provide more effective feature enhancement weights by iteratively learning from previous actions aim to achieve higher prediction confidence, further promoting the classification accuracy.

Ablation Study. We first conduct ablation experiments to validate the design of our proposed different components. We compare the following different settings. (1) Baseline: the 3D network described in Section 2.3 without agent modules. We train the model with $Loss = L_{O_{T1}} + L_{O_{T1ce}} + L_{O_{T2}} + L_{O_{FLAIR}} + 4 \times L_{O_{fusion}}$, and the prediction is produced as $Outputs = O_{T1} + O_{T1ce} + O_{T2} + O_{FLAIR} + O_{fusion}$. (2) Ours w/o Inter: the proposed method with RL-based intra-learning, while the modality-specific features are concatenated without inter-learning. (3) Ours (1 modal): the proposed method using one modality, *i.e.* T1, without inter-learning. (4) Ours (2 modal): the proposed method using two modalities, *i.e.* T1 and FLAIR. (5) Ours (3 modal): the proposed method using three modalities, *i.e.* T1, T2 and FLAIR. (6) Ours: the proposed method with RL-based intra-/inter- modality learning using four modalities.

Table 3. Evaluations (mean \pm standard deviation) of proposed methods on LNDb.

| Method | Acc | Sen | Spe | F1 score |
|----------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Baseline | 0.718 \pm 0.007 | 0.507 \pm 0.013 | 0.480 \pm 0.006 | 0.473 \pm 0.006 |
| 3D CBAM [16] | 0.732 \pm 0.027 | 0.496 \pm 0.019 | 0.498 \pm 0.021 | 0.490 \pm 0.020 |
| Baseline w/ ia | 0.744\pm0.012 | 0.520\pm0.019 | 0.502\pm0.010 | 0.496\pm0.010 |

Table 2 shows the ablation results. It is observed that the proposed method improves the performance of classification by adopting the proposed RL-based intra-/inter- modality learning. Especially, RL-based intra-modality learning contributes an accuracy improvement of 5% as the input 3D volumes are enhanced by the RL-based proposed weights, while the RL-based inter-modality enhancement gets an improvement of 1.1% in accuracy on this dataset. It is worth noting that our method has better performance than other attention-based fusion methods (compared to 3D CBAM [16] and 3D MFB [18]), which demonstrates that our design of RL-based modality enhancement is better. In addition, from the results, it can be seen that the prediction performance of our method improves when using more modalities, which also verifies the effectiveness of multi-modality learning. We adopt the T1 and FLAIR for Ours (2 modal) and 3D MFB [18] to demonstrate that our method can get better performance using the same modalities.

Validation of the RL-based Enhancement. To demonstrate the effectiveness of our method, we further explore the performance of RL-based intra-modality learning on LNDb [12] which contains a total of 229 lung nodule CT images from the training set and distinguish lung nodules into three texture classes (solid, sub-solid, and GGO). For experiments of LNDb, we first extract $96 \times 96 \times 96$ cubes from the whole CT scans according to the given center location of a lung nodule. Only one agent is used for intra-modality learning in the LNDb classification model. The batch size is set to 4 and we adopt a 5-fold cross-validation strategy for performance evaluation.

As shown in Table 3, RL-based enhancement also promotes the classification accuracy of LNDb by about 2.6% compared to the baseline, which demonstrates the efficiency of our method on another imaging data. Moreover, we also compare the performance of CBAM [16] on the intra-modality learning on the LNDb, proving that the proposed method helps the model to learn better intra-modal representations with higher metrics.

4 Conclusion

This paper innovatively introduces the RL strategy into the intra-modality learning and inter-modality learning and proposes a novel hierarchic feature enhancement framework for multi-modality learning. With the purpose of exploiting complementary inter-modality features while preserving intra-modality features, the multi-modality learning problem is modeled as a dynamic hierarchic feature en-

hancement issue. In addition, the proposed RL-based multi-modality learning method is general and has great potential to boost the performance of various medical image tasks. Our future work will focus on different more effective training strategies and extend our framework to other multi-modality medical image analysis problems.

Acknowledgement. This work was supported by the National Key Research and Development Program of China (2019YFE0113900).

References

1. Browning, J., Kornreich, M., Chow, A., Pawar, J., Zhang, L., Herzog, R., Odry, B.L.: Uncertainty aware deep reinforcement learning for anatomical landmark detection in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 636–644. Springer (2021)
2. Fan, J., Cao, X., Yap, P.T., Shen, D.: Birnet: Brain image registration using dual-supervised fully convolutional networks. *Medical image analysis* **54**, 193–206 (2019)
3. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. *Neural computation* **16**(12), 2639–2664 (2004)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
5. Khvostikov, A., Aderghal, K., Benois-Pineau, J., Krylov, A., Catheline, G.: 3d cnn-based classification using smri and md-dti images for alzheimer disease studies. arXiv preprint arXiv:1801.05968 (2018)
6. Lei, W., Su, Q., Gu, R., Wang, N., Liu, X., Wang, G., Zhang, X., Zhang, S.: One-shot weakly-supervised segmentation in medical images. arXiv preprint arXiv:2111.10773 (2021)
7. Lin, Y.Y., Liu, T.L., Fuh, C.S.: Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(6), 1147–1160 (2010)
8. Morani, K., Unay, D.: Deep learning based automated covid-19 classification from computed tomography images. arXiv preprint arXiv:2111.11191 (2021)
9. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
10. Nie, D., Lu, J., Zhang, H., Adeli, E., Wang, J., Yu, Z., Liu, L., Wang, Q., Wu, J., Shen, D.: Multi-channel 3d deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Scientific reports* **9**(1), 1–14 (2019)
11. Nie, D., Zhang, H., Adeli, E., Liu, L., Shen, D.: 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In: International conference on medical image computing and computer-assisted intervention. pp. 212–220. Springer (2016)
12. Pedrosa, J., Aresta, G., Ferreira, C., Rodrigues, M., Leitão, P., Carvalho, A.S., Rebelo, J., Negrão, E., Ramos, I., Cunha, A., et al.: Lndb: a lung nodule database on computed tomography. arXiv preprint arXiv:1911.08434 (2019)
13. Peng, Y., Huang, X., Qi, J.: Cross-media shared representation by hierarchical learning with multiple deep networks. In: IJCAI. pp. 3846–3853 (2016)

14. Saha, A., Tushar, F.I., Faryna, K., D'Anniballe, V.M., Hou, R., Mazurowski, M.A., Rubin, G.D., Lo, J.Y.: Weakly supervised 3d classification of chest ct using aggregated multi-resolution deep segmentation features. In: *Medical Imaging 2020: Computer-Aided Diagnosis*. vol. 11314, p. 1131408. International Society for Optics and Photonics (2020)
15. Wang, A., Lu, J., Cai, J., Cham, T.J., Wang, G.: Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Transactions on Multimedia* **17**(11), 1887–1898 (2015)
16. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 3–19 (2018)
17. Zhou, F., Li, T., Li, H., Zhu, H.: Tpcnn: two-phase patch-based convolutional neural network for automatic brain tumor segmentation and survival prediction. In: *International MICCAI Brainlesion Workshop*. pp. 274–286. Springer (2017)
18. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE transactions on medical imaging* **39**(9), 2772–2781 (2020)
19. Zhou, T., Fu, H., Zhang, Y., Zhang, C., Lu, X., Shen, J., Shao, L.: Multi-modal multi-channel network for overall survival time prediction of brain tumor patients. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 221–231. Springer (2020)
20. Zhou, T., Thung, K.H., Zhu, X., Shen, D.: Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human brain mapping* **40**(3), 1001–1016 (2019)