Variational multimodal distillation for diagnosing plaque vulnerability in carotid 3D MRI

Bo Cao[1], Mengmeng Feng[1], Fan Yu[1], Zhen Qian[2], and Jie Lu[1]

[1] Department of Radiology and Nuclear Medicine, Xuanwu Hospital, Capital Medical University, China
[2] Beijing United Intelligent Imaging Research Institute, China

Abstract. Multimodal learning has attracted much attention in recent years due to its ability to effectively utilize data features from a variety of different modalities. Diagnosing the vulnerability of atherosclerotic plaques directly from carotid 3D MRI images is relatively challenging for both radiologists and conventional 3D vision networks. In clinical practice, radiologists assess patient conditions using a multimodal approach that incorporates various imaging modalities and domain-specific expertise, paving the way for the creation of multimodal diagnostic networks. In this paper, we have developed an effective strategy to leverage radiologists' domain knowledge to automate the diagnosis of carotid plaque vulnerability through **V**ariation inference and **M**ultimodal knowledge **D**istillation (**VMD**). This method excels in harnessing cross-modality prior knowledge from limited image annotations and radiology reports within training data, thereby enhancing the diagnostic network's accuracy for unannotated 3D MRI images. We conducted in-depth experiments on the dataset collected in-house and verified the effectiveness of the VMD strategy we proposed. Code will be available at this url.

Keywords: Multimodal learning · Knowledge distillation · Atherosclerosis plaque · 3D MRI classification.

1 Introduction

The rupture of atherosclerotic plaques in the carotid artery is a significant factor in ischemic stroke events [15]. Plaque vulnerability is determined by multiple factors such as thin-cap fibroatheroma (TCFA), lipid-rich necrotic core (LRNC), inflammatory cell infiltration, and intraplaque hemorrhage (IPH) [3]. Although fully-supervised learning models are capable of automating the identification of these components, the detailed manual annotation of arterial wall boundaries and plaque components involves a considerable amount of time. Reducing the scope of annotation to exclude plaque components may reduce the time required by two-thirds, but this simplification impedes the ability to directly assess plaque vulnerability through segmentation outcomes [20,4]. Therefore, it becomes highly appealing to enable neural networks to learn from limited anatomical structure annotations and achieve more precise automated diagnosis of plaque vulnerability in unannotated carotid 3D MRI images.

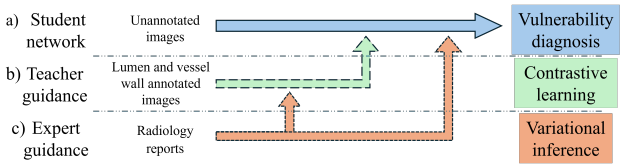

Fig. 1: **How limited annotations and reports work in VMD.** The diagnosis of plaque vulnerability is directly performed using unannotated 3D MRI images by the student network (a). The limited annotations and radiology reports, which act as teacher (b) and expert (c) respectively, influence the student by maximizing contrastive learning-based MI and variational inference. This process significantly enhances the diagnostic capabilities of the student.

Contrastive learning-based mutual information (MI) optimization excels in multimodal learning [13,17,7], enabling effective cross-modal feature transfer and alignment, similar to knowledge distillation (KD) [8]. Limited annotated data, focusing on the lumen and vessel wall, provide noise-reduced anatomical information for diagnosing plaque vulnerability and could be considered an additional modality for multimodal learning. Exploiting the similarity between annotated and unannotated data, we speculate that samples of the same category, whether they are vulnerable or stable plaques, from these two modalities should exhibit higher degree of similarity. Therefore, we have developed a contrastive learning-based KD network aimed at improving the diagnostic accuracy for unannotated images. This is achieved by maximizing MI between the student model, which lacks annotations, and the teacher model, which has access to limited annotations. This approach is designed to leverage the inherent similarities within the data to enhance the learning process, enabling the student model to better identify and diagnose plaque vulnerability without the need for extensive annotated data, as shown in Fig. 1(b).

Radiology reports offer a deep insight into images as interpreted by radiologists, encapsulating high-level domain knowledge. This expertise can be integrated into the aforementioned networks, offering expert-level guidance to both the teacher and student models. Therefore, we have designed an additional expert network to improve the previous distillation strategy for both the teacher and student networks. Some studies have confirmed that pre-trained biomedical language models can effectively comprehend medical texts [22,21,9] and transfer domain knowledge from text to 2D medical image diagnostic networks in multimodal learning [16,19]. Among them, the KD method based on variational inference does not require additional costs at the inference stage [19], making it suitable for direct diagnosis from unannotated images. This approach, which minimizes a specific Kullback-Leibler Divergence (KLD) to estimate the evidence lower bound (ELBO) of the optimization objective, facilitates the transfer of knowledge across different modalities and can be considered as another way of maximizing MI. Considering the higher complexity of 3D images compared to

2D images, in the expert guidance network, we employ variational inference and contrastive learning to maximize MI between the image networks and the expert network, which allows the teacher network to retain the domain knowledge transfer from the expert network as much as possible while guiding the student.

In this study, we developed a novel three-level distillation strategy based on **V**ariation inference and **M**ultimodal knowledge **D**istillation(**VMD**) to transfer comprehensive understanding of radiology reports and limited annotations to student network for plaque vulnerability diagnosis by a student-teacher-expert structure, which enabled us to accomplish the complex task at a lower cost. Overall, our contributions are summarized as follows:

1. The student-teacher-expert distillation strategy VMD effectively transfers the prior knowledge from radiology reports and the limited annotations into the problem of diagnosing plaque vulnerability from unannotated 3D images, which performs best in the main evaluation metrics.
2. By introducing limited homologous annotations with a student-teacher distillation structure, we have expanded the application of contrastive learning-based MI optimization in medical knowledge transfer and solved more challenging problems at a lower cost.
3. We imposed a tighter constraint on the teacher network integrating variational inference based on MI optimization, thereby more effectively utilizing the rich knowledge from radiology reports to enhance the diagnostic performance of the student network.

2 Related Work

Student-Teacher distillation structure The concept of KD was first proposed in [8] and is considered highly effective in model compression and feature transfer. The student-teacher distillation structure used therein has been applied in many complex tasks. VID [2] maximizes the MI between the student and the teacher network by minimizing the entropy of the teacher to effectively transfer knowledge to the student. CRD [17] introduces contrastive learning into the student-teacher structure. CCL [7] separates image and audio in video classification tasks and optimizes the performance through feature mixing contrast learning. And this idea is further explained in AMID [5] from the perspective of MI among different modalities. However, our method, i.e., VMD, utilizes manual annotations to acquire anatomical knowledge from the unannotated images and feed it back to the student by maximizing inter-class MI.

Image-Text Medical Knowledge Transfer Pre-trained biomedical semantic understanding models have demonstrated commendable proficiency in comprehending medical texts in recent years [21,9,22]. Some studies [19,18,24] guided the neural network in diagnosing diseases in 2D image modality by extracting features closely associated with the diagnosis from text by a pre-trained language model. These methods employ either knowledge distillation [16,8] or contrastive

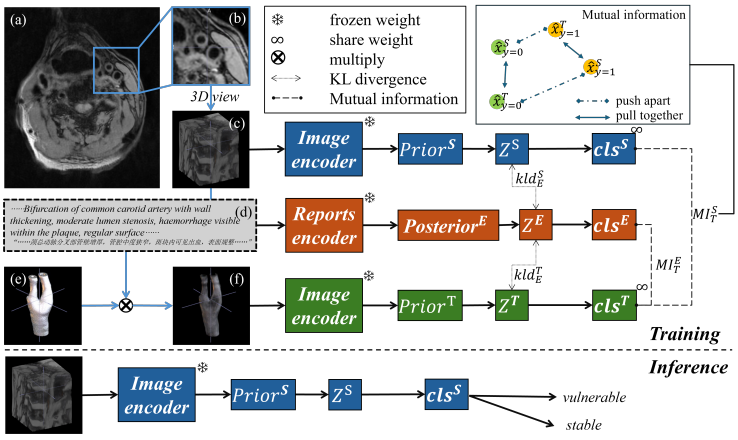

Fig. 2: **The overview of VMD.** (a) is the transverse section of the original 3D MRI image. The carotid artery region (b-c) for the student network is obtained through cropping. The input (f) for the teacher network is the product of (c) and the lumen and vessel wall annotation (e). The radiology report (d) corresponding to the 3D MRI is for the expert network. During the training phase, we maximize the MI and minimize the KL divergence between the latent spaces z of different networks.

learning [14] to align features in order to address the problem. By integrating contrastive learning with conditional variational inference, we expanded the strategy of VKD [19] into a student-teacher-expert structure to provide a tighter constraint to transfer the knowledge from the expert to teacher networks because of the substantial disparity between 3D and 2D images.

3 Methodology

The overall architecture of VMD is illustrated in Fig. 2. In order to efficiently transfer cross-modal prior knowledge from the Expert network E (dark orange) and teacher network T (green) to the weaker-performing student network S (blue), we need to maximize the MI between different student-teacher pairs [17]. During the optimization process, we also minimize the cross-entropy H_{cls} of all networks as much as possible to make the prediction approach the true value. The global optimization objective o is shown in Eq. 1, α_1,α_2,α_3,α_4 are weights for the corresponding terms.

$$o = \alpha_1 \mathcal{I}(S,T) + \alpha_2 \mathcal{I}(S,E) + \alpha_3 \mathcal{I}(T,E) - \alpha_4 H_{cls} \tag{1}$$

where the $\mathcal{I}$ means MI. We will sequentially solve for each term of the equation in the following sections.

Maximizing mutual information based on comparative learning: $\mathcal{I}(S,T)$ For the homologous annotated x^T and unannotated x^S, suppose a pair of predictions $\hat{x}_j^T$ and $\hat{x}_i^S$ derived from T and S, we aim to minimize the distance between predictions with the same class label and maximize the gap between predictions with different classes for all N samples. To maximize the MI between S and T, we consider applying a similarity-based model $M\left(\hat{x}_j^T, \hat{x}_i^S\right)$ inspired by [5] to quantify the difference between $\hat{x}_j^T$ and $\hat{x}_i^S$, shown in Eq. 2.

$$M\left(\hat{x}_j^T, \hat{x}_i^S\right) = \frac{\exp\left(\phi \frac{\hat{x}_j^T, \hat{x}_i^S}{\tau}\right)}{\sum_{k=1}^{N} \exp\left(\phi \frac{\left(\hat{x}_j^T, \hat{x}_k^S\right)}{\tau}\right)} \tag{2}$$

where ϕ is a cosine similarity scoring function, and temperature τ is set empirically to 0.5. For B samples of the same category out of all N samples, we choose to optimize the $\mathcal{I}(S,T)$ in a form that looks like infoNCE loss [13] but has additional supervision guided by class label, as shown in Eq. 3,

$$\mathcal{I}(S,T) = \frac{1}{N} \sum_{j=1}^{N} \left[\sum_{i=1}^{B} \log M\left(\hat{x}_j^T, \hat{x}_i^S\right) + \sum_{k=1}^{N-B} \log\left(1 - M\left(\hat{x}_j^T, \hat{x}_k^S\right)\right) \right] \tag{3}$$

Cross-modal knowledge transfer from reports: $\mathcal{I}(S,E)$ Conditional variational inference [16] is highly suitable for structured prediction tasks, and VKD [19] has validated its effectiveness in understanding medical image analysis tasks. We construct a variational probability model and formulate the problem of predicting the target y from the unannotated image x^S as solving the posterior probability $p\left(y|x_I^S\right)$. In order to obtain the following conditional likelihood formula, we introduce the latent variable z^S:

$$\log p\left(y|x_I^S\right) = \log \int p\left(y|x_I^S, z_I^S\right) p\left(z_I^S|x_I^S\right) dz_I^S \tag{4}$$

the $p\left(z_I^S|x_I^S\right)$is the conditional prior for x^S, which is easy to obtain by neural network. The prediction problem can be translated into maximizing this integral w.r.t. z. Thus, we need to find the maximized posterior probability $p\left(z^S|x^S, y\right)$, which is relatively difficult to solve. Inspired by [19], we introduce the representation z^E from the expert network and use the variational posterior distribution $q\left(z^E|x^E\right)$ to approximate $p\left(z^S|x^S, y\right)$. Given that the KLD is well behaved in the distance metrics [11], we can obtain kld_E^S in Fig. 2 by applying Bayesian formulae given the weight θ of the networks, and translating the original problem

into minimizing the KLD between $q\left(z^E|z^E\right)$ and $p\left(z^S|x^S\right)$:

$$\begin{aligned} kld_E^S &= D_{KL}\left[q_\theta\left(z^E|x^E\right)||p_\theta\left(z^S|x^S,y\right)\right] \\ &= \log p_\theta\left(y|x^S\right) + \int q_\theta\left(z^E|x^E\right)\log\frac{q_\theta\left(z^E|x^E\right)}{p_\theta\left(z^S|x^S,y\right)p_\theta\left(z^S,x^S\right)}d\theta \\ &= \log p_\theta\left(y|x^S\right) - \mathbb{E}_{q_\theta(z^E|x^E)}\left[\log p_\theta\left(y|x^S,z^S\right)\right] \\ &+ D_{KL}\left[q_\theta\left(z^E|x^E\right)||p_\theta\left(z^E|x^S\right)\right] \end{aligned} \tag{5}$$

from where we can maximize the $\mathcal{I}\left(S,E\right)$ by maximizing the $ELBO_{SE}$ as follows:

$$\mathcal{I}\left(S,E\right) = ELBO_{SE} = \mathbb{E}_{q(z^E|x^E)}\left[\log p\left(y|x^S,z^S\right)\right] - D_{KL}\left[q\left(z^E|x^E\right)||p\left(z^E|x^S\right)\right] \tag{6}$$

The further constraint on the teacher: $\mathcal{I}\left(T,E\right)$ The further constraint on the teacher stems from both variational inference and contrastive learning. Therefore, we first compute the KLD between the latent spaces z based on Eq. 6 and then compute the similarity between the predictions based on Eq. 3. It is noteworthy that T and S share the same parameters in the classification head. Eq. 7 shows the summary of $\mathcal{I}\left(T,E\right)$.

$$\begin{aligned} \mathcal{I}\left(T,E\right) &= \mathbb{E}_{q(z^E|x^E)}\left[\log p\left(y|x^T,z^T\right)\right] - D_{KL}\left[q\left(z^E|x^E\right)||p\left(z^E|x^T\right)\right] \\ &+ \frac{1}{N}\sum_{j=1}^{N}\left[\sum_{i=1}^{B}\log M\left(\hat{x}_j^T,\hat{x}_i^E\right) + \sum_{k=1}^{N-B}\log\left(1-M\left(\hat{x}_j^T,\hat{x}_k^E\right)\right)\right] \end{aligned} \tag{7}$$

At last, we use the minimized cross-entropy loss function $\mathcal{L}_{ce}$ to further obtain more accurate classifications for all 3 parts of VMD, as shown in Eq. 8, λ_1,λ_2,λ_3 are different weights for these terms.

$$H_{cls} = \lambda_1\mathcal{L}_{ce}\left(\hat{x}_i^T,y_i\right) + \lambda_2\mathcal{L}_{ce}\left(\hat{x}_i^S,y_i\right) + \lambda_3\mathcal{L}_{ce}\left(\hat{x}_i^E,y_i\right) \tag{8}$$

4 Experiments

Dataset Upon exclusion of images that were unreported, postoperative, of poor quality, or depicted occluded vessels, we included 502 3D-T1-FSE sequence im-ages of unilateral vessels from 303 participants who met the diagnosis of carotid stenosis following the guidelines [1]. The scanning center is positioned at carotid bifurcation, enabling us to crop the image into a volume of 128×128×60. Each vessel image was labeled as "vulnerable" or "stable" by two experienced radi-ologists, resulting in a total of 350 vulnerable 3D MRI images and 152 stable images. We divided the dataset into training, validation, and test sets in a ra-tio of 4:1:1. During the training phase, we used five-fold cross-validation and employed a class-weighted random sampling strategy to avoid class imbalance issues.

Implementation We use pre-trained and frozen weight encoders for network inputs. Considering the richness of 3D image information, training a 3D MRI visual feature extraction network from scratch is not cost-effective. Therefore, we first employ MedicalNet [6] and chose to freeze the network weights pre-trained on 23 3D CT and MRI datasets and fine-tune using a two-layer MLP network. Then we apply amortization techniques and reparameterization [11] to get the prior $p(z_I|x_I)$ for both T and S. We use a two-layer MLP whose latent feature size is 512 to play the role of z. For the expert network E, we apply Mc-BERT [21], a Chinese medical semantic understanding BERT-base network pre-trained in a considerably large Chinese biomedical text corpus [23], as the report encoder and use one single transformer layer to fine-tune it. Using techniques akin to those for image input, we compute the posterior $q\left(z^E|x^E\right)$. All experiments

Table 1: **Compare classification performance with other methods**, **bold** represnets the best, underline represents the second best.

Setting	ROC	Accuracy	Precision	Recall	Fbeta
Baseline[6]	0.6567±0.0018	0.6588±0.0007	0.7802±0.0055	0.7213±0.0063	0.7487±0.003
VPT[10]	0.6141±0.0244	0.6195±0.0055	0.7044±0.0074	0.7966±0.0077	0.7476±0.0009
VID[2]	0.6594±0.018	0.7171±0.0055	0.7959±0.0016	**0.8069±0.0077**	**0.8014±0.0046**
VKD[19]	0.6818±0.0162	0.678±0.0185	0.8065±0.0203	0.7172±0.0094	0.7592±0.0123
CCL[7]	0.659±0.0005	0.6361±0.0044	0.778±0.0045	0.6779±0.0123	0.7238±0.0035
VMD	**0.7136±0.013**	**0.7244±0.0204**	**0.8197±0.0192**	0.7828±0.0094	0.8008±0.0135

were performed in two NVIDIA A40 GPUs, using the Adam optimizer for 400 epochs. The learning rate is initialized to 5e-4 and weight decay is 1e-4. Five random seeds were used to calculate the average and standard deviation of ROC and other evaluation metrics. The β of Fbeta score is 0.5.

5 Results

We compare VMD with the baseline method, three SOTA knowledge distillation methods, and a network fine-tuning method, namely VID [2], VKD [19], CCL [7], and VPT [10]. We employed the fully fine-tuned 3DResNet-50 network as our classification baseline, a model derived from MedicalNet [6].

Compare with SOTA methods VPT [10] facilitates network fine-tuning by introducing few learnable parameters into the original 2D images, but the advantage is not evident in 3D images. VID [2] minimizes the entropy of the teacher network to transfer knowledge from the teacher to the student. VKD [19] transfers abundant knowledge from the Electronic Health Record (EHR) to the X-ray diagnosis network by introducing variational knowledge distillation. CCL [7] introduces a contrastive learning-based MI optimization method in the cross-modal knowledge distillation problem, realizing feature transfer from audio and images to video. For the sake of fairness in comparative experiments, we used the

fine-tuned baseline method 3DResNet-50 [6] as the image encoder for different methods to compare. Table 1 shows the comparison results of the classification performance of VMD and the other SOTA methods in unannotated carotid artery 3D MRI images. Thanks to the diagnostic knowledge derived from cross-modal variational inference and the anatomical prior gained by maximizing inter-class MI, our method achieves the best in most evaluation metrics and is only slightly different from the best results in Recall and Fbeta.

Table 2: **Ablation studies for our distillation strategy**.

(a) ablation study for student w/o teacher (Eq. 3,7) and expert (Eq. 6,7) to diagnose unannotated images.

Setting	ROC	Accuracy	Precision	Recall	Fbeta
w/o both	0.6645±0.0193	0.6537±0.0109	0.7859±0.0181	0.7034±0.0463	0.7412±0.0197
w/o teacher	0.6818±0.0162	0.678±0.0185	0.8065±0.0203	0.7172±0.0094	0.7592±0.0123
w/o expert	0.6815±0.028	0.6512±0.0164	0.76±0.0175	0.7421±0.0016	0.7505±0.0087
VMD	**0.7109±0.0163**	**0.722±0.0181**	**0.8191±0.0186**	**0.7793±0.0077**	**0.7987±0.0115**

(b) ablation study for teacher w/o expert (Eq. 7) to diagnose annotated images.

Setting	ROC	Accuracy	Precision	Recall	Fbeta
w/o expert	0.7372±0.0048	0.6732±0.0134	0.8151±0.0101	0.6966±0.0378	0.7505±0.018
teacher	**0.7923±0.0316**	**0.7829±0.0181**	**0.8222±0.0487**	**0.7945±0.0229**	**0.8079±0.0344**

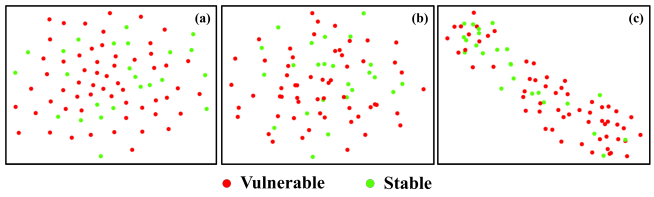

Fig. 3: **Visualization of z by t-SNE[12]**, (**a**) student without guidance; (**b**) student with teacher's guidance only; (**c**) student with both two guidances.

Ablation study To validate the roles of teacher network and expert network in VMD, we designed two ablation experiments for the teacher network and the student network respectively. In Table 2(a), when applying both the teacher network and the expert network, we achieve the best performance in diagnosing unannotated 3D MRI images from the student network. Table 2(b) substantiates that when we eliminate the interfering signals surrounding the carotid artery by introducing annotations, the components and positional information from the radiology report can guide the classification network more efficiently. We also

use t-SNE [12] to visualize the latent features z for classification, as in Fig. 3, the distribution of features with two guidance is more separated for the two classes.

6 Conclusion

The task of automated diagnosis of vulnerable plaques in carotid 3D MRI images is challenging. In this study, we propose a student-teacher-expert distillation strategy named VMD for the automated diagnosis network of carotid 3D MRI images, which achieves a significant improvement in diagnosing the vulnerability of atherosclerotic plaques. By combining variational inference and contrastive learning-based MI optimization theory, we efficiently transfer the cross-modal information in the teacher network and the expert network to the student network, without introducing additional inference costs.

Acknowledgments. This study was funded by National Natural Science Foundation of China (No. 82130058) and National Natural Science Foundation of China (No. 81974261).

References

1. AbuRahma, A.F., Avgerinos, E.D., Chang, R.W., Darling III, R.C., Duncan, A.A., Forbes, T.L., Malas, M.B., Murad, M.H., Perler, B.A., Powell, R.J., et al.: Society for vascular surgery clinical practice guidelines for management of extracranial cerebrovascular disease. Journal of vascular surgery **75**(1), 4S–22S (2022)
2. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9155–9163 (2019). https://doi.org/10.1109/CVPR.2019.00938
3. Cai, J.M., Hatsukami, T.S., Ferguson, M.S., Small, R., Polissar, N.L., Yuan, C.: Classification of human carotid atherosclerotic lesions with in vivo multicontrast magnetic resonance imaging. Circulation **106**(11), 1368–1373 (2002)
4. Chen, L., Zhao, H., Jiang, H., Balu, N., Geleri, D.B., Chu, B., Watase, H., Zhao, X., Li, R., Xu, J., et al.: Domain adaptive and fully automated carotid artery atherosclerotic lesion detection using an artificial intelligence approach (latte) on 3d mri. Magnetic Resonance in Medicine **86**(3), 1662–1673 (2021)
5. Chen, M., Xing, L., Wang, Y., Zhang, Y.: Enhanced multimodal representation learning with cross-modal kd. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11766–11775 (June 2023)
6. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
7. Chen, Y., Xian, Y., Koepke, A., Shan, Y., Akata, Z.: Distilling audio-visual knowledge by compositional contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7016–7025 (2021)
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

9. Huang, K., Altosaar, J., Ranganath, R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
10. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision (ECCV) (2022)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
13. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
15. Saba, L., Saam, T., Jäger, H.R., Yuan, C., Hatsukami, T.S., Saloner, D., Wasserman, B.A., Bonati, L.H., Wintermark, M.: Imaging biomarkers of vulnerable carotid plaques for stroke risk prediction and their potential clinical implications. The Lancet Neurology **18**(6), 559–572 (2019)
16. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)
17. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
18. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
19. Van Sonsbeek, T., Zhen, X., Worring, M., Shao, L.: Variational knowledge distillation for disease classification in chest x-rays. In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27. pp. 334–345. Springer (2021)
20. Wu, J., Xin, J., Yang, X., Sun, J., Xu, D., Zheng, N., Yuan, C.: Deep morphology aided diagnosis network for segmentation of carotid artery vessel wall and diagnosis of carotid atherosclerosis on black-blood vessel wall mri. Medical physics **46**(12), 5544–5561 (2019)
21. Xu, Z., Gong, L., Ke, G., He, D., Zheng, S., Wang, L., Bian, J., Liu, T.Y.: Mc-bert: Efficient language pre-training via a meta controller. arXiv preprint arXiv:2006.05744 (2020)
22. Yasunaga, M., Leskovec, J., Liang, P.: Linkbert: Pretraining language models with document links. In: Association for Computational Linguistics (ACL) (2022)
23. Zhang, N., Jia, Q., Yin, K., Dong, L., Gao, F., Hua, N.: Conceptualized representation learning for chinese biomedical text mining. arXiv preprint arXiv:2008.10813 (2020)
24. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: MICCAI (2023)