From Slices to Volumes: Multi-Scale Fusion of 2D and 3D Features for CT Scan Report Generation

Abdullah Hosseini[0000−0003−2967−3033], Ahmed Ibrahim[0009−0002−9217−4173], and Ahmed Serag[0000−0002−4145−5509]

AI Innovation Lab, Weill Cornell Medicine-Qatar
afs4002@qatar-med.cornell.edu

Abstract. The increasing complexity of medical imaging data underscores the necessity for multimodal intelligent systems capable of integrating diverse data representations for comprehensive and precise analysis. In the domain of 3D CT scans, the generation of accurate and clinically meaningful medical reports requires both volumetric contextual information and the fine-grained spatial details inherent in 2D slices. To address this challenge, we propose a framework that employs a pretrained 2D self-supervised learning encoder, initially trained on CT scan slices integrated with a 3D aggregator. By combining the rich, high-resolution information from 2D slices with the spatial coherence of 3D volumetric data, our approach maximizes the complementary strengths of both representations. Experimental results demonstrate that our method outperforms existing baseline approaches in both report generation and multiple-choice question answering, highlighting the critical role of multidimensional feature integration. This work underscores the transformative potential of multimodal intelligent systems in bridging complex imaging data with practical clinical insights, ultimately improving radiological diagnostics and patient care[1].

Keywords: 3D Medical Imaging · Medical Report Generation · Multimodal Large Language Model · Self Supervised Learning.

1 Introduction

Generating Reports and performing Visual Question-Answering (VQA) for three dimensional Computed Tomography (CT) scans is highly specialized and time consuming, requiring the delivery of precise, and easily comprehensible medical information. Thanks to recent advances in Artificial Intelligence (AI), particularly Large Language Models (LLMs) [6,9], multimodal medical imaging has undergone a significant transformation in both 2D and 3D medical image analysis [17,18]. This progress is especially evident in 3D medical image report generation, where extracting key insights from volumetric data is crucial for accurate diagnosis and treatment planning.

[1] Our code is now available at github.com/serag-ai/SAMF

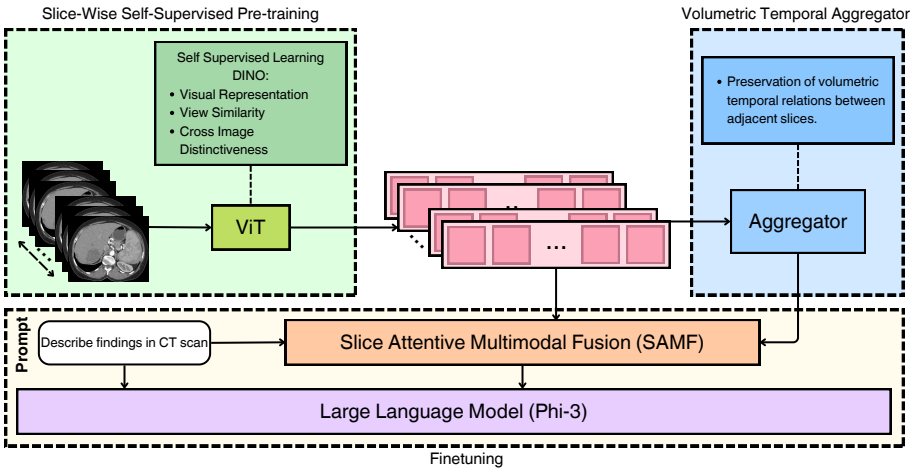

Fig. 1. Architectural overview of the proposed medical report generation framework, incorporating slice-wise encoding, user prompts, and volumetric features through SAMF fusion methodology.

However, most existing models for 3D medical report generation are trained exclusively on either 3D or 2D data, which are often limited in size and resolution, overlooking the rich information available in 2D slices of CT scans [1,11,5,10]. In contrast, radiologists typically adopt a dual-perspective approach, combining global and local viewpoints. The **global perspective**, akin to a 3D encoder, captures high-level spatial features, allowing radiologists to understand the overall structural topology of the volumetric data. The **local perspective**, similar to a 2D encoder, focuses on individual slice planes, emphasizing fine-grained details and textural nuances.

To address this challenge, we propose a methodology that utilize both the high-level spatial features of 3D data and the rich local details of 2D slices. Our approach begins by pretraining a 2D encoder using a self-supervised learning(SSL) framework (e.g., DINO [4]) on CT scan slices from three planes: axial, coronal, and sagittal. The outputs of this 2D encoder are then processed by a 3D aggregator to preserve volumetric temporal relationships between slices. Additionally, we introduce a novel fusion technique that integrates the outputs of the aggregator, the 2D encoder, and a prompt, effectively bridging the gap between 2D and 3D representations (see Figure 1). This fused representation is then fed into an LLM to generate medical reports.

Our proposed method enhances both medical report generation and multiple-choice question answering accuracy in clinical applications by combining high-resolution 2D features with 3D volumetric spatial relationships. Our key contributions are: (1) A comprehensive radiology report generation framework tailored for 3D medical imaging, integrating both 2D and 3D data representations. (2) A novel fusion technique, Slice-Attentive Multimodal Fusion (SAMF), designed

to seamlessly combine 2D and 3D features, enabling richer and more contextually aware representations. (3) Open-sourcing our trained models and code to promote reproducibility and facilitate further research.

2 Methodology

Existing studies primarily focus on encoding either full 3D CT volumes or isolated 2D slices, with limited research on synergistically fusing both 2D and 3D representations for radiology report generation and VQA tasks. Section 2.1 outlines our approach for feature extraction from volumetric data. Section 2.2 presents our key contribution, a novel fusion mechanism that systematically integrates embeddings from the 2D encoder, 3D volumetric aggregator, and user-specified prompts. Section 2.3 details the publicly available datasets used in this study.

2.1 CT Scan Feature Extraction

For pre-training our 2D slice encoder, we employ the state-of-the-art SSL framework DINO, encoding 2D slices extracted from three orthogonal planes: axial, coronal, and sagittal. We adopt a Vision Transformer(ViT) [8] as the backbone, processing slice images ($s \in S$) from 3D CT scans (X_{ct}), at a resolution of 224x224 pixels. The input images are tokenized into a sequence of fixed-size 16x16 pixel patches, which are then linearly embedded for efficient processing within the transformer architecture.

These extracted 2D slice-based features serve as inputs to a 3D aggregator, which processes the embeddings produced by the 2D encoder. Let f_{2d} and f_{3d} represent the 2D encoder and 3D aggregator, respectively. The feature extraction process is formulated as follows:

$$\boldsymbol{z}_{\mathrm{2d}}^{\boldsymbol{s}} = f_{2d}\left(s\right), \forall s \in S \tag{1}$$

$$\boldsymbol{z}_{\mathrm{3d}} = f_{3d}\left(\{\boldsymbol{z}_{\mathrm{2d}}^{\boldsymbol{s}}\}_{s \in S}\right) \tag{2}$$

As an ablation study, prior to final fine-tuning of the LLM, the 3d aggregator f_{3d} is warmed-up in a self-supervised manner to preserve the volumetric temporal relations between adjacent slices' embeddings. We posited that implementing a warm-up phase for the aggregator prior to final fine-tuning could improve model performance by mitigating the effects of random weight initialization and enabling the aggregator to acquire preliminary feature representations from CT scan data. During this stage, the 2D encoder f_{2d} remains frozen, while the 3D aggregator is optimized using a contrastive loss function. Specifically, the self-supervised pre-training is performed by maximizing the mutual information between embeddings of adjacent slices from different planes of a CT scan using InfoNCE loss [15].

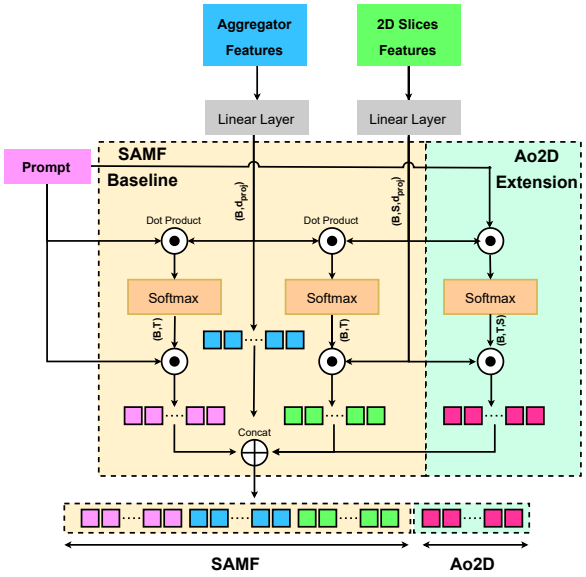

Fig. 2. Slice-Attentive Multi-Modal Fusion architecture. The framework projects three feature spaces (z_{3d}, text tokens, and z^s_{2d}) into a shared representation space for cross-modal fusion and interaction.

2.2 Slice-Attentive Multi-Modal Fusion (SAMF)

Inspired by radiologists' dual-perspective diagnostic mechanism, we introduce SAMF, a fusion strategy that integrates information from three distinct feature spaces: the 3D aggregator, textual tokens, and 2D slice embeddings. In SAMF, we first project the input features, denoted as z_{3d} and z_{2d}, into a unified embedding space of dimension d_{proj}. This projection ensures that features derived from 3D, 2D, and textual modalities are aligned within a common embedding space of consistent dimensionality.

$$\boldsymbol{F}_{2d}^{\text{proj}} = \mathbf{W}_{2d} \cdot z_{\text{2d}}^{s}, \quad \in \mathbb{R}^{B\times S\times d_{\text{proj}}} \tag{3}$$

$$\boldsymbol{F}_{3d}^{\text{proj}} = \mathbf{W}_{3d} \cdot z_{\text{3d}}, \quad \in \mathbb{R}^{B\times d_{\text{proj}}} \tag{4}$$

Similarly, we utilize attention mechanisms to capture the relationships between 3D features, 2D slices, and text tokens. This approach determines the relevance of 2D slices, text tokens, and 3D feature to each other. The attention mechanism is implemented as follows:

$$\boldsymbol{\alpha}_{\text{2d3d}} = \text{softmax}\left(\mathbf{F}_{\text{2d}}^{\text{proj}} \cdot \left(\mathbf{F}_{\text{3d}}^{\text{proj}}\right)^{\top}\right), \quad \in \mathbb{R}^{B\times S} \tag{5}$$

$$\boldsymbol{\beta}_{\text{t3d}} = \text{softmax}\left(\mathbf{F}_{\text{text}}^{\text{proj}} \cdot \left(\mathbf{F}_{\text{3d}}^{\text{proj}}\right)^{\top}\right), \quad \in \mathbb{R}^{B\times T} \tag{6}$$

$$\boldsymbol{\gamma}_{\text{t2d}} = \text{softmax}\left(\mathbf{F}_{\text{text}}^{\text{proj}} \cdot \left(\mathbf{F}_{\text{2d}}^{\text{proj}}\right)^{\top}, \text{dim} = S\right), \quad \in \mathbb{R}^{B\times T\times S} \tag{7}$$

Softmax is applied to normalize the attention scores into weights, and these weights are used to aggregate the input features. This mechanism ensures that the model prioritizes the most informative components of the 2D, 3D, and textual data, recognizing that not all slices or tokens hold equal relevance to one another.

$$\boldsymbol{F}_{\text{2d3d}}^{\text{agg}} = \alpha^{T} \cdot \mathbf{F}_{2d}^{\text{proj}}, \quad \in \mathbb{R}^{B\times 1\times d_{\text{proj}}} \tag{8}$$

$$\boldsymbol{F}_{\text{text3d}}^{\text{agg}} = \beta^{T} \cdot \mathbf{F}_{3d}^{\text{proj}}, \quad \in \mathbb{R}^{B\times 1\times d_{\text{proj}}} \tag{9}$$

At the final stage fusion startegy, we expand them and concatenate them to form the final fused representation:

$$\mathbf{F}_{\text{fused}} = \mathbf{W}_{\text{fusion}} \cdot \left[\ \mathbf{F}_{3d}^{\text{proj}} \ || \ \mathbf{F}_{\text{2d3d}}^{\text{agg}} \ || \ \mathbf{F}_{\text{text3d}}^{\text{agg}} \ \right] \tag{10}$$

During the fine-tuning stage, $\mathbf{F}_{\text{fused}}$ serves as the input to the LLM, specifically Phi-3, for generating the final radiology report. Our fusion strategy is designed to be flexible, allowing integration of additional modalities or features as needed.

Attention Over 2D Slices (Ao2D) To further refine SAMF, we introduce Attention Over 2D Slices (Ao2D), which captures fine-grained relationships between text tokens and 2D slices. This enhancement aggregates text-to-2D features via averaging over text tokens:

$$\mathbf{F}_{\text{text2d}}^{\text{agg}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{S}\boldsymbol{\gamma}_{\text{t2d},t,s}\mathbf{F}_{\text{2d},s}^{proj}, \quad \in \mathbb{R}^{B\times 1\times d_{\text{proj}}} \tag{11}$$

Finally, these features are concatenated to construct the final fused representation:

$$\mathbf{F}_{\text{fused}} = \mathbf{W}_{\text{fusion}} \cdot \left[\ \mathbf{F}_{3d}^{\text{proj}} \ \| \ \mathbf{F}_{\text{2d3d}}^{\text{agg}} \ \| \ \mathbf{F}_{\text{text3d}}^{\text{agg}} \ \| \ \mathbf{F}_{\text{text2d}}^{\text{agg}} \ \right] \tag{12}$$

2.3 Dataset

Our study utilized different publicly available datasets tailored to specific tasks. For the self-supervised pretraining of the 2D encoder, we employed the publicly available Liver Tumor Segmentation (LiTS) dataset [3], along with an additional publicly available dataset for organ localization [19]. These datasets were accessed via the MedMnist-v2 framework [20] and were used across all three anatomical planes: axial, coronal, and sagittal. Each view consisted of approximately 59,000, 24,000, and 25,000 slices, respectively.

For the LLM fine-tuning stage, we utilized the publicly available CT-RATE dataset [11], which includes approximately 50,000 post-processed CT scans paired with corresponding radiology reports and multiple-choice questions. Model performance was evaluated using the same dataset split provided in CT-RATE, ensuring a direct comparison with the baseline results reported in their work.

3 Experiments

3.1 Evaluation Metrics

To assess the effectiveness of incorporating SAMF in generating radiology reports, we employed a combination of natural language generation (NLG) metrics and LLM-based evaluations. The NLG metrics included BLEU-1 and BLEU-4 [16], METEOR [2], ROUGE-L [14], and the BERT-F1 score [21]. These metrics respectively measure lexical similarity, synonym-aware matching, structural coherence, and contextual semantic similarity by using BERT embeddings to evaluate meaning beyond exact word overlap.

For LLM-based evaluations, we used the LLAMA-70B model to assess the clinical relevance of the generated medical reports. Following the approach used for evaluating CT-CHAT [10], we applied the same prompting strategy to categorize Llama-generated scores for each question into three distinct ranges. For the VQA task, specifically multiple-choice question answering, we evaluated performance using accuracy, precision, and recall of correct answer prediction.

3.2 Results and Comparison with Baselines

To evaluate the efficacy of our proposed SAMF framework and its extension Ao2D, we conducted comparative analyses against state-of-the-art methods on the CT-Rate benchmark. Since the CT-Rate dataset is relatively new, most relevant studies are recent and available on arXiv, making the reproducibility of their results challenging. Therefore, we reported their values directly from their papers. For CT2Rep [11], due to the inaccessibility of pretrained weights, we retrained the model using their best-reported parameters.

The results of our experiments are presented in Table 1, showcasing the performance of *SAMF + Ao2D* in comparison to existing baselines. Our approach achieves either the best or second best performance across all metrics, highlighting the effectiveness of the SAMF and Ao2D in improving the quality of radiology report generation. Figure 3A demonstrates an example of a report generated by both methods.

For multiple-choice question answering, the results are presented in Table 2, with additional insights into detailed predictions available in Figure 3B. The findings show that *SAMF + Ao2D* outperforms SAMF alone and the baseline CT-Chat across all evaluated metrics, demonstrating the effectiveness of slice-attentive multimodal fusion in enhancing multiple-choice question answering for chest CT analysis. It should be noted that, the fine-tuning stage required approximately 16 hours on a single NVIDIA A100 GPU equipped with 80 GB of RAM.

Table 1. Comparative analysis of our approach against state-of-the-art methods for medical report generation for chest CT-scans. **Bold** values indicate best performance, while underlined values represent second-best results.

Model	Bleu1	Bleu4	RougeL	Meteor	Bert F1	Llama Score
E3D-GPT [12]	0.412	-	-	**0.418**	0.880	-
MS-VLM [13]	-	0.232	**0.438**	0.396	-	-
CT-AGRG [7]	-	0.172	0.280	0.196	0.867	-
CT2Rep [11]	0.309	0.172	0.243	0.173	0.865	6.35
CT-Chat [10]	0.395	-	0.321	0.219	-	5.664
Our Baseline (SAMF)	0.423	0.203	0.338	0.356	0.879	6.792
+ *Ao2D*	**0.440**	**0.261**	0.417	0.417	**0.889**	**7.165**

Table 2. Comparative analysis of SAMF and Ao2D against state-of-the-art methods for multiple-choice question answering for chest CT-scans.

Model	Bleu1	RougeL	Meteor	Llama Score	Accuracy	Precision	Recall	F1
CT-Chat [10]	0.838	0.895	0.577	0.901	-	-	-	-
SAMF	0.940	0.925	0.926	0.939	0.915	0.898	0.893	0.895
+ *Ao2D*	**0.942**	**0.929**	**0.930**	**0.942**	**0.920**	**0.902**	**0.900**	**0.901**

3.3 Ablation Study

We conducted ablation studies by (1) selectively freezing different components of our framework during fine-tuning and (2) implementing a warming-up stage for the aggregator through a self-supervised approach before fine-tuning stage. The purpose of freezing different components was twofold: first to isolate the informational contribution of each component to the final prediction, and second to empirically evaluate its individual roles in the pipeline.

As demonstrated in Table 3, the fusion approach incorporating Ao2D highly sensitive to component freezing, leading to diminished performance. This sensitivity can be attributed to Ao2D's inherent reliance on the 2D embedding space derived from the image encoder for pattern recognition. Its architectural design,

Table 3. Results for the medical report generation task of selectively freezing different components of the framework and implementing a self-supervised warming-up stage for the aggregator.

Model	ViT	Aggregator	BLEU-1	BLEU-4	ROUGE-L	METEOR
SAMF baseline	✓		0.426	0.209	0.345	0.366
+ *SSL Warmup*			0.442	0.184	0.377	0.412
	✓		0.447	0.186	0.300	0.375
	✓	✓	0.339	0.128	0.282	0.308
+ *Ao2D*	✓		0.257	0.089	0.260	0.263
+ *Ao2D, SSL Warmup*			0.445	0.204	0.338	0.365
	✓	✓	0.223	0.419	0.091	0.255

which focuses on attention over two-dimensional features, makes it dependent on maintaining an adaptable 2D encoder.

Fig. 3. A) Qualitative evaluation of medical report generation. Correct predictions are highlighted, while incorrect predictions are indicated in red font. B) Comparison of model performance of SAMF and Ao2D on multiple-choice question answering.

4 Discussion

In this study, we introduced Slice-Attentive Multimodal Fusion (SAMF), an innovative approach for radiology report generation from 3D chest CT volumes. Our method leverages transfer learning, transitioning from slice-wise encoding to volume-level feature extraction through a SSL paradigm. The effectiveness of our framework is further enhanced by Attention over 2D slices (Ao2D) on top of SAMF, which captures fine-grained relationships between text tokens and 2D slices. This is particularly useful for tasks that require detailed alignment between language and visual data.

Our methodology demonstrates improved fidelity in radiological report generation and multiple-choice question answering, effectively bridging the gap between visual features and textual descriptions. Although our results could be further improved by incorporating additional clinically focused evaluation metrics, our successful integration of both 2D and 3D representations marks a significant advancement in multimodal AI for medical imaging and clinical decision support.

Acknowledgments. The authors gratefully acknowledge the support of the IT and administration teams at Weill Cornell Medicine-Qatar for facilitating the computational and operational aspects of this study.

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

References

1. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
5. Chen, H., Zhao, W., Li, Y., Zhong, T., Wang, Y., Shang, Y., Guo, L., Han, J., Liu, T., Liu, J., et al.: 3d-ct-gpt: Generating 3d radiology reports through integration of large vision-language models. arXiv preprint arXiv:2409.19330 (2024)
6. DeepSeek-AI, D.G., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
7. Di Piazza, T.: Ct-agrg: Automated abnormality-guided report generation from 3d chest ct volumes. arXiv preprint arXiv:2408.11965 (2024)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
10. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Durugol, O.F., Wittmann, B., Amiranashvili, T., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography (2024)
11. Hamamci, I.E., Er, S., Menze, B.: Ct2rep: Automated radiology report generation for 3d medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–486. Springer (2024)
12. Lai, H., Jiang, Z., Yao, Q., Wang, R., He, Z., Tao, X., Wei, W., Lv, W., Zhou, S.K.: E3d-gpt: Enhanced 3d visual foundation for medical vision-language model. arXiv preprint arXiv:2410.14200 (2024)
13. Lee, C., Park, S., Shin, C.I., Choi, W.H., Park, H.J., Lee, J.E., Ye, J.C.: Read like a radiologist: Efficient vision-language model for 3d medical imaging interpretation. arXiv preprint arXiv:2412.13558 (2024)

14. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
15. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4004–4012 (2016)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
17. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
18. Tang, Y., Yuan, Y., Tao, F., Tang, M.: Cross-modal augmented transformer for automated medical report generation. IEEE Journal of Translational Engineering in Health and Medicine (2025)
19. Xu, X., Zhou, F., Liu, B., Fu, D., Bai, X.: Efficient multiple organ localization in ct image using 3d region proposal network. IEEE transactions on medical imaging **38**(8), 1885–1898 (2019)
20. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
21. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)