CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging

Ibrahim Ethem Hamamci[1], Sezgin Er[2], and Bjoern Menze[1]

[1]University of Zurich   [2]Istanbul Medipol University
{ibrahim.hamamci@uzh.ch}

Abstract. Medical imaging plays a crucial role in diagnosis, with radiology reports serving as vital documentation. Automating report generation has emerged as a critical need to alleviate the workload of radiologists. While machine learning has facilitated report generation for 2D medical imaging, extending this to 3D has been unexplored due to computational complexity and data scarcity. We introduce the first method to generate radiology reports for 3D medical imaging, specifically targeting chest CT volumes. Given the absence of comparable methods, we establish a baseline using an advanced 3D vision encoder in medical imaging to demonstrate our method's effectiveness, which leverages a novel autoregressive causal transformer. Furthermore, recognizing the benefits of leveraging information from previous visits, we augment CT2Rep with a cross-attention-based multi-modal fusion module and hierarchical memory, enabling the incorporation of longitudinal multimodal data. Access our code at https://github.com/ibrahimethemhamamci/CT2Rep.

Keywords: 3D Medical Imaging, Chest CT Volume, Radiology Report, CT-RATE Dataset, Report Generation, Longitudinal, Transformers

1 Introduction

The integration of machine learning into radiology, driven by numerous public datasets [12, 14, 28], has significantly enhanced disease classification and segmentation [6, 10, 25, 30]. Furthermore, recent advancements have enabled the development of many methods for generating radiology reports for 2D medical imaging [4,15,20,26,27] utilizing public datasets [16,23]. However, this progress in report generation has not yet extended to 3D medical imaging, due to computational complexities [8] and the lack of datasets paired with radiology reports [3].

3D medical imaging, such as computed tomography (CT) and magnetic resonance imaging, provides a more detailed perspective on the patient's condition compared to 2D imaging [22]. Consequently, manual report generation, essential for conveying diagnostic findings, becomes more time-consuming and error-prone, highlighting the need for automation. One of the challenges in developing such a framework lies in the scarcity of 3D medical imaging datasets paired with reports [19]. Moreover, the nature of 3D images involves volumetric data, which necessitates more sophisticated algorithms for interpreting the additional

dimension. This complexity presents unique obstacles to generating descriptive and clinically relevant reports that effectively capture the details of 3D images.

Recognizing this gap, our work introduces CT2Rep, the first approach to automated radiology report generation for 3D medical imaging, specifically targeting chest CT volumes. CT2Rep leverages a novel 3D auto-regressive causal vision feature extractor, optimized for processing volumetric data. We also incorporate relational memory to utilize information from previous report generations, employing memory-driven conditional layer normalization to integrate this data into our framework. To train our framework, we utilize the CT-RATE dataset [9], which consists of 25,692 non-contrast chest CT volumes, expanded to 50,188 through various reconstructions, from 21,304 unique patients, along with corresponding radiology reports. CT2Rep's uniqueness, the first of its kind in 3D medical imaging, means that no directly comparable methods exist. Nonetheless, to demonstrate the effectiveness of our framework, we reasonably designed a baseline using a state-of-the-art vision encoder used for 3D chest CT volume interpretation, CT-Net [6], for report generation. CT2Rep outperforms this well-designed baseline method, showcasing the efficacy of our novel approach.

Radiologists typically assess a 3D chest CT volume alongside previous volumes and reports for the same patient, as multiple visits are common in clinical practice. Longitudinal volumes and their reports contain valuable information, and leveraging this multimodal data can potentially enhance report generation. Hence, we extended CT2Rep by incorporating a cross-attention-based multimodal fusion module coupled with a hierarchical memory-driven decoder. This extension not only addresses computational challenges associated with 3D image analysis but also facilitates the inclusion of longitudinal multimodal patient data, enriching the context and accuracy of generated reports. We evaluated this extended version, named CT2RepLong, through a comprehensive ablation study to underscore the importance of historical imaging and reports in informing current diagnostic interpretations. Our contributions can be summarized as:

- We propose CT2Rep, the first radiology report generation framework for 3D medical imaging, employing a novel auto-regressive causal transformer.
- As CT2Rep is the first of its kind and no comparable methods exist, we have designed a baseline employing the cutting-edge 3D vision encoder used in chest CT classification to benchmark our method and prove its effectiveness.
- We augment CT2Rep with a cross-attention-based multi-modal fusion module and a hierarchical memory-driven decoder to leverage commonly available longitudinal data, backed by a comprehensive ablation study showcasing the efficacy of incorporating longitudinal data for report generation.
- We make our trained models and source codes publicly available to facilitate out-of-the-box report generation for 3D chest CT volumes.

2   Methods

Although 3D medical imaging, such as 3D chest CT volumes, offers more comprehensive information than its 2D counterparts like chest X-rays, there are

currently no solutions for generating radiology reports for 3D imaging due to data scarcity and computational complexity. To address this gap, we developed a 3D sequence-to-sequence generation model, detailed in Sec. 2.1, utilizing the data outlined in Sec. 2.3. Additionally, we enhanced our method to incorporate longitudinal multimodal data from previous visits, as described in Sec. 2.2.

2.1   The Proposed Method

The model (Φ_{CT2Rep}) accepts input 3D volumes $x \in \mathbb{R}^{(240) \times 480 \times 480}$ as a sequence of CT patches $x = \{x_1, x_2, x_3, ..., x_N\}, x_n \in \mathbb{R}^{(12) \times 24 \times 24}$ to predict a target sequence $r_{out} = \{r_1^{\mathrm{out}}, r_2^{\mathrm{out}}, ..., r_T^{\mathrm{out}}\}, r_t^{\mathrm{out}} \in \mathbb{V}$. Here, N represents CT feature count, T the token count, and $\mathbb{V}$ the possible token vocabulary. CT2Rep, depicted in Fig. 1, consists of three key components, each elaborated below.
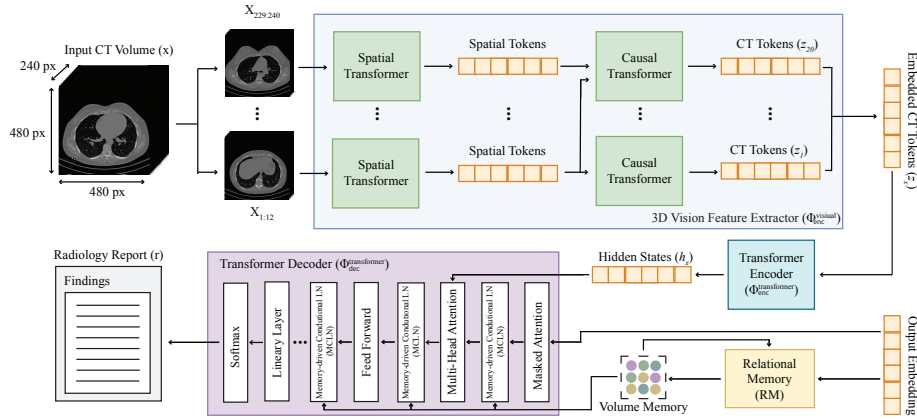

Fig. 1: CT2Rep features a novel auto-regressive causal transformer for 3D vision feature extraction, complemented by RM and MCLN-enhanced transformer-based encoder and decoder network for clinically accurate report generation.

3D vision feature extractor. A key component and main contribution of our framework, this network ($\Phi_{\mathrm{enc}}^{\mathrm{visual}}$) facilitates the extraction of embedded CT tokens from 3D chest CT volumes by segmenting the data into distinct patches and transforming them into a lower-dimensional latent space, inspired by [1]. These tokens capture essential information, facilitating subsequent analysis.

The network takes a 3D CT volume (x) and produces embedded CT tokens $z_x \in \mathbb{R}^{20 \times 20 \times 20 \times 512}$, by initially extracting $(12) \times 24 \times 24$ non-overlapping patches from x. Each patch is then mapped to a D-dimensional space, with D set to 512. The patches are then reshaped and transformed linearly to $B \times T \times \frac{H}{p_1} \times \frac{W}{p_2} \times D$, following a previous work [11]. Here, p_t denotes the temporal patch size, T represents the number of temporal patches, B is the batch size, H and W are the height and width of the slices, respectively, and p_1 and p_2 represent the

spatial patch sizes. After patch embedding, the resulting tensor size is $B \times (T) \times \frac{H}{p_1} \times \frac{W}{p_2} \times D$. This tensor is then processed by two transformer networks consecutively. First, the spatial transformer operates on a reshaped tensor of size $(B \cdot (T)) \times (\frac{H}{p_1} \cdot \frac{W}{p_2}) \times D$, yielding a tensor of the same dimensions. Subsequently, the causal transformer processes this output reshaped to $(\frac{H}{p_1} \cdot \frac{W}{p_2}) \times (B \cdot (T)) \times D$, and produces an output maintaining these dimensions. This method ensures that both spatial and latent dimensions are preserved after each layer, thereby retaining 3D volumetric information throughout the network's processing stages. The overall 3D chest CT volume feature extraction process, formally defined as $z_x = \Phi_{\mathrm{enc}}^{\mathrm{visual}}(x)$, ensures 3D volumetric information is preserved, facilitating the effective construction of sequence-to-sequence models for report generation.

Transformer encoder. We employ a conventional transformer ($\Phi_{\mathrm{enc}}^{\mathrm{transformer}}$) to encode CT features extracted by $\Phi_{\mathrm{enc}}^{\mathrm{visual}}$. This network processes these features to produce encoded hidden states via an attention mechanism, crucial for capturing feature interdependencies. The encoded hidden states are represented as:

$$h_x = \{h_1, h_2, \ldots, h_N\} = \Phi_{\mathrm{enc}}^{\mathrm{transformer}}(z_x) = \Phi_{\mathrm{enc}}^{\mathrm{transformer}}(z_1, z_2, \ldots, z_N),$$

where each $h_n \in \mathbb{R}^{512}$ represents the encoded state of a patch, with N being the total patch count. The attention mechanism in the transformer is defined as $\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where Q, K, and V stand for the query, key, and value matrices, respectively, and d_k is the key's dimensionality.

Transformer decoder. We adapt a traditional transformer network as a decoder ($\Phi_{\mathrm{dec}}^{\mathrm{transformer}}$), with two notable enhancements. First, we integrate relational memory (RM) [4], entailing the utilization of a matrix to encapsulate and propagate pattern information across generation steps. Each row within this matrix stores specific pattern details, which are iteratively refined through updates incorporating outputs from preceding steps. The updating mechanism involves employing the matrix from the previous step as a query and concatenating it with the prior output to serve as the key and value for the transformer's multi-head attention module. Mathematically, this process is achieved through multi-head attention, where $Q = M_{t-1} \cdot W_q$, $K = [M_{t-1}; y_{t-1}] \cdot W_k$, and $V = [M_{t-1}; y_{t-1}] \cdot W_v$. Here, y_{t-1} denotes the embedding of the previous step's output, while W_q, W_k, and W_v represent the trainable weights for query, key, and value transformations, respectively. Thus, the model effectively learns conserved report patterns, such as *"Trachea, both main bronchi are open."*, within similar CT volumes. Second, we employ a memory-driven conditional layer normalization (MCLN) [17], integrating RM directly into the decoder's scaling (γ) and shifting (β) parameters. This makes the model more contextually aware and adept at generating accurate text outputs. The decoding process is defined as:

$$r_T^{\mathrm{out}} = \Phi_{\mathrm{dec}}^{\mathrm{transformer}}(h_1, h_2, \ldots, h_N, MCLN(RM(r_1^{\mathrm{out}}, r_2^{\mathrm{out}}, \ldots, r_{T-2}^{\mathrm{out}}, r_{T-1}^{\mathrm{out}}))).$$

Inference. After training, CT2Rep (Φ_{CT2Rep}) is able to generate a radiology report (r_{out}) for a given 3D chest CT volume (x), formally defined as follows:

$$r_{\mathrm{out}} = \Phi_{\mathrm{CT2Rep}}(x) = \Phi_{\mathrm{dec}}^{\mathrm{transformer}}(\Phi_{\mathrm{enc}}^{\mathrm{transformer}}(\Phi_{\mathrm{enc}}^{\mathrm{visual}}(x))).$$

2.2   Longitudinal Data Utilization

To utilize multimodal data from previous visits, we augmented CT2Rep with a cross-attention-based fusion module [31] that allows to predict an outcome sequence $r_{out}^{new} = \{r_1, r_2, ..., r_T\}, r_t \in \mathbb{V}$, for a given new 3D chest CT volume (x^{new}) by integrating representations from the previous CT volume (x^{old}) and its corresponding previous report (r_{in}^{old}). The fusion process is facilitated by computing cross-attention between previous volume and report representations by $R^* = \text{softmax}\left(\frac{q(H_{RP})k(H_{IP})^\top}{\sqrt{d_k}}\right)$ and $I^* = \text{softmax}\left(\frac{q(H_{IP})k(H_{RP})^\top}{\sqrt{d_k}}\right)$, where H_{IP} and H_{RP} are the attended features for longitudinal volumes and reports, respectively. These features are concatenated to create a comprehensive multimodal longitudinal representation H_L. This integrated approach significantly enhances the performance of the longitudinal framework, $\Phi_{\text{CT2RepLong}}$, by leveraging both spatial and semantic information from previous visits, as detailed in Fig. 2.
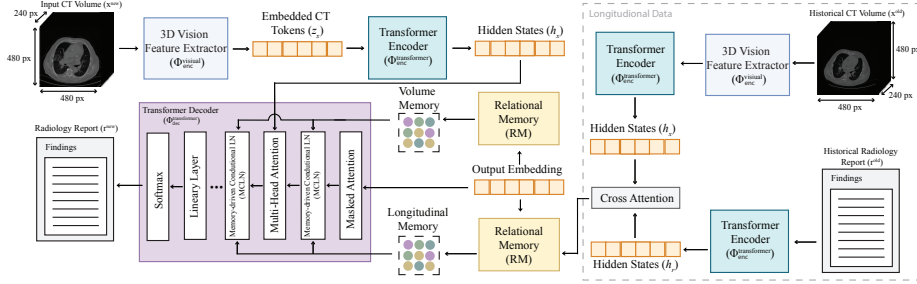

Fig. 2: CT2RepLong enhances CT2Rep with a cross-attention multi-modal fusion module and longitudinal memory for effective historical data integration.

Multimodal transformer decoder. The decoder of $\Phi_{\text{CT2RepLong}}$ closely follows that of Φ_{CT2Rep}. However, $\Phi_{\text{CT2RepLong}}$ employs two more cross-attention mechanisms, together defined as $\Phi_{\text{long}}^{\text{attn}}$, to analyze the relationships between previous reports (r^{old}) and volumes (x^{old}), and vice versa. The outputs are concatenated, and then RM is applied to the new report as per the process described in Sec. 2.1. Subsequently, another cross-attention, $(\Phi_{\text{mem}}^{\text{attn}})$, is used between the RM and the cross-attention outputs from previous volumes and reports. The resulting cross-attention outputs are then utilized in MCLN, formalized as:

$$r_T^{\text{out}} = \Phi_{\text{dec}}^{\text{transformer}}(h_1, ..., h_N, MCLN(\Phi_{\text{mem}}^{\text{attn}}(RM(r_1^{\text{out}}, ..., r_{T-1}^{\text{out}}), \Phi_{\text{long}}^{\text{attn}}(r^{\text{old}}, x^{\text{old}})))).$$

Inference. After training, $\Phi_{\text{CT2RepLong}}$ can generate a report $(r_{\text{out}}^{\text{new}})$ for a given new volume (x^{new}), alongside a previous volume and its corresponding report:

$$r_{\text{out}}^{\text{new}} = \Phi_{\text{CT2RepLong}}(x^{\text{new}}, x^{\text{old}}, r^{\text{old}})$$
$$= \Phi_{\text{dec}}^{\text{transformer}}(\Phi_{\text{enc}}^{\text{transformer}}(\Phi_{\text{enc}}^{\text{visual}}(x^{\text{new}})), \Phi_{\text{long}}^{\text{attn}}(r^{\text{old}}, x^{\text{old}})).$$

2.3   Dataset Preparation.

We utilize 3D chest CT volumes along with corresponding radiology reports from the publicly available CT-RATE dataset [9]. For the development of CT2Rep, we employ all volumes and reports from the initial release of CT-RATE. Our dataset comprises 25,701 non-contrast 3D chest CT volumes from 21,314 unique patients, which expands to 49,138 volumes after applying multiple reconstructions tailored to different window settings [29]. Each volume features a resolution of 512×512 pixels in the axial plane, with slice counts ranging from 100 to 600. The radiology reports associated with each volume are segmented into four sections: clinical information, technique, findings, and impression; however, only the findings section is utilized for report generation training. The same radiology report is used for each reconstructed volume of a single CT volume. The dataset is divided into a training set of 20,000 patients and a validation set of 1,314 patients, ensuring no overlap. CT volumes were converted to Hounsfield Units (HU) using slope and intercept values from the metadata and clipped to $[-1000 \text{ HU}, +200 \text{ HU}]$ to represent the practical diagnostic limits of the HU scale [5]. Each volume was subsequently resized to achieve uniform spacing of 0.75 mm on the x and y axes and 1.5 mm on the z-axis. The volumes were either center-cropped or padded to achieve a consistent resolution of $(240) \times 480 \times 480$. **Creating the longitudinal dataset.** We targeted patients with more than two visits, yielding 6,766 and 429 3D chest CT volumes from 2,638 and 169 unique patients for the training and validation sets, respectively. After applying various reconstructions, these volumes increased to 13,354 for training and 849 for validation. We chronologically ordered the volumes for each patient using the *StudyTime* metadata attribute and paired every two possible longitudinal volumes for a patient, resulting in 28,441 training and 1,689 validation pairs.

3   Experiments and Results

To evaluate model efficacy in generating radiology reports, we employed natural language generation (NLG) and clinical efficacy (CE) metrics. NLG metrics include BLEU (BL) [24], METEOR (M) [18], and ROUGE-L (R_L) [21], assessing word overlap, synonym use and word order, and sequence matching, respectively. For CE metrics, we fine-tuned the CXR-Bert model [2] for multi-label classification of reports on 18 abnormalities, as detailed in the supplementary material. We then predicted the abnormality labels of both ground-truth data and generated reports and computed classification scores, including precision (P), recall (R), and F1 score, to measure the clinical accuracy of the generated reports.

3.1   Comparison with the Baseline Method

Given the absence of directly comparable methods, further highlighting our method's novelty, we established a benchmark for radiology report generation by implementing a state-of-the-art vision encoder, CT-Net [7], used in 3D medical imaging. CT-Net is the first and only model developed for classifying 3D

Table 1: Quantitative evaluation showcases the effectiveness of our CT2Rep model in comparison to a well-designed baseline and highlights how our enhanced method, CT2RepLong, leverages longitudinal data to improve performance.

Method	NLG Metrics						CE Metrics		
	BL-1	BL-2	BL-3	BL-4	M	R_L	P	R	F1
Base w/ CT-Net	0.443	0.399	0.375	0.354	0.286	0.442	0.513	0.531	0.456
CT2Rep (Ours)	**0.460**	**0.415**	**0.390**	**0.369**	**0.295**	**0.459**	**0.749**	**0.548**	**0.534**
methods below utilize longitudinal data									
Baseline	0.372	0.317	0.282	0.251	0.238	**0.353**	0.666	0.465	0.525
+ report	0.330	0.284	0.260	0.241	0.213	0.313	0.623	0.410	0.524
+ volume	0.305	0.261	0.238	0.220	0.204	0.291	0.662	0.434	0.530
+ report + volume	0.365	0.319	0.292	0.271	0.239	0.351	0.658	0.410	0.533
CT2RepLong (Ours)	**0.374**	**0.327**	**0.304**	**0.401**	**0.285**	0.263	**0.727**	**0.511**	**0.536**

chest CT volumes. Its architecture comprises a ResNet-18 feature extractor [13], augmented by 3D convolutional blocks designed to streamline ResNet features, followed by final classification layers. In our approach, we harness the feature extraction capabilities of CT-Net, using these features as inputs for our 3D volume transformer, establishing it as the baseline for our study. Table 1 demonstrates that our CT2Rep significantly outperforms this baseline, thanks to our novel auto-regressive causal transformer used as the 3D vision feature extractor.

Case study. We assessed our model's performance through a qualitative analysis on a randomly chosen case from our test set, comparing generated reports to the ground truth. Figure 3 illustrates that CT2Rep accurately generates reports with content flow and medical terminology that closely resemble those written by radiologists, markedly surpassing the baseline established with CT-Net.
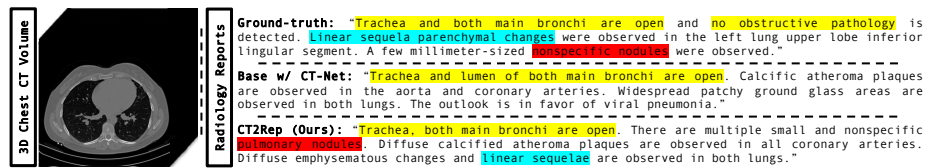

Fig. 3: Comparison of ground-truth with reports generated by a CT-Net-based baseline and CT2Rep, highlighting CT2Rep's medical precision with color codes.

3.2 Ablation Study on Longitudinal Data Utilization

We evaluated CT2RepLong's performance and the impact of incorporating prior data through an ablation study. Initially, we established a baseline by training

CT2Rep solely on the longitudinal dataset without using any prior data. We then augmented this baseline with three strategies: utilizing embeddings from previous reports, previous volume embeddings, and their combination via simple fusion (excluding our longitudinal cross-attention mechanism). Table 1 demonstrates the advantages of prior multimodal data and our unique cross-attention mechanism. The exception with R_L can be attributed to its emphasis on sequence length over the enriched content and diversity from longitudinal data integration. Besides, despite the limited size of longitudinal data—only 13% of patients (see Sec. 2.3)—CT2RepLong's performance was comparable with the original CT2Rep, illustrating its effectiveness, even with a constrained dataset.

Case study. A qualitative analysis of a random test case (Fig. 4) reveals that CT2RepLong significantly benefits from integrating longitudinal data. Key terms like "cardiomegaly" and "calcified atherosclerotic plaques" appeared in both current and previous reports, enriching the accuracy of the generated reports. Notably, terms missed by the baseline, such as "cardiomegaly" were included by CT2RepLong, aligning with the ground truth and demonstrating the enhanced reliability of report generation with our extension for utilizing longitudinal data.
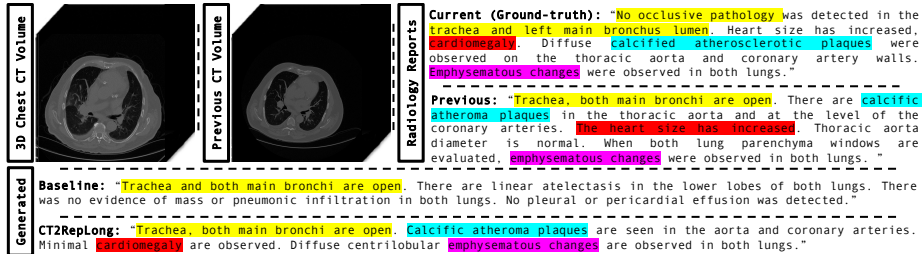

Fig. 4: CT2RepLong surpasses the baseline, leveraging longitudinal data for enhanced medical detail accuracy, with related terms color-coded for clarity.

3.3 Implementation Details

CT2Rep and the baseline method (see Sec. 3.1) were trained on 49,138 3D CT volumes and their corresponding reports (Sec. 2.3). We used the Adam optimizer with β_1 and β_2 hyperparameters set to 0.9 and 0.99, respectively. The learning rate was established at 0.00005 for the visual extractor and 0.0001 for the other parameters. A StepLR scheduler with a gamma of 0.1, a batch size of 1, and a maximum token count of 300 for the scheduler were employed. CT2RepLong and the ablation methods (Sec. 3.2) were trained on 28,441 pairs (Sec. 2.3), utilizing the same hyperparameters as CT2Rep. The training duration for all models was one week on a single NVIDIA A100 GPU, achieving 20 epochs. Inference takes approximately 35 seconds for CT2Rep and 50 seconds for CT2RepLong.

4   Discussion and Conclusion

In conclusion, we introduce CT2Rep, the first framework for automating 3D medical imaging report generation, with a focus on chest CT volumes. Leveraging an innovative auto-regressive causal transformer architecture and integrating relational memory, CT2Rep enhances accuracy in report generation. As the first of its kind, we establish a benchmark using the state-of-the-art vision encoder in 3D chest CT volume interpretation to showcase CT2Rep's effectiveness. Additionally, we extend its capabilities with longitudinal data integration, resulting in CT2RepLong, further enhancing context and accuracy. We make our trained models and code fully open-source to lay a solid foundation for further research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846 (2021)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
3. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79**, 102444 (2022)
4. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
5. DenOtter, T.D., Schubert, J.: Hounsfield unit (2019)
6. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical image analysis **67**, 101857 (2021)
7. Draelos, R.L., Dov, D., Mazurowski, M.A., Lo, J.Y., Henao, R., Rubin, G.D., Carin, L.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical image analysis **67**, 101857 (2021)
8. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems **35**, 31841–31854 (2022)
9. Hamamci, I.E., Er, S., Almas, F., Simsek, A.G., Esirgun, S.N., Dogan, I., Dasdelen, M.F., Wittmann, B., Simsar, E., Simsar, M., et al.: A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. arXiv preprint arXiv:2403.17834 (2024)

10. Hamamci, I.E., Er, S., Simsar, E., Sekuboyina, A., Gundogar, M., Stadlinger, B., Mehl, A., Menze, B.: Diffusion-based hierarchical multi-label object detection to analyze panoramic dental x-rays. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 389–399. Springer Nature Switzerland, Cham (2023)

11. Hamamci, I.E., Er, S., Simsar, E., Tezcan, A., Simsek, A.G., Almas, F., Esirgun, S.N., Reynaud, H., Pati, S., Bluethgen, C., et al.: Generatect: Text-guided 3d chest ct generation. arXiv preprint arXiv:2305.16037 (2023)

12. Hamamci, I.E., Er, S., Simsar, E., Yuksel, A.E., Gultekin, S., Ozdemir, S.D., Yang, K., Li, H.B., Pati, S., Stadlinger, B., et al.: Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays. arXiv preprint arXiv:2305.19112 (2023)

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

14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)

15. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)

16. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

17. Lample, G., Sablayrolles, A., Ranzato, M., Denoyer, L., Jégou, H.: Large memory layers with product keys. Advances in Neural Information Processing Systems **32** (2019)

18. Lavie, A., Denkowski, M.J.: The meteor metric for automatic evaluation of machine translation. Machine translation **23**, 105–115 (2009)

19. Li, J., Zhu, G., Hua, C., Feng, M., Bennamoun, B., Li, P., Lu, X., Song, J., Shen, P., Xu, X., et al.: A systematic collection of medical image datasets for deep learning. ACM Computing Surveys **56**(5), 1–51 (2023)

20. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)

21. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)

22. Müller, N.: Computed tomography and magnetic resonance imaging: past, present and future. European Respiratory Journal **19**(35 suppl), 3s–12s (2002)

23. Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)

24. 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)

25. Pati, S., Thakur, S.P., Hamamcı, İ.E., Baid, U., Baheti, B., Bhalerao, M., Güley, O., Mouchtaris, S., Lang, D., Thermos, S., et al.: Gandlf: the generally nuanced deep

learning framework for scalable end-to-end clinical workflows. Communications Engineering **2**(1), 23 (2023)

26. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. Nature medicine **29**(8), 1930–1940 (2023)

27. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision. pp. 563–579. Springer (2022)

28. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)

29. Willemink, M.J., Noël, P.B.: The evolution of image reconstruction for ct—from filtered back projection to artificial intelligence. European radiology **29**, 2185–2195 (2019)

30. Yüksel, A.E., Gültekin, S., Simsar, E., Özdemir, S.D., Gündogar, M., Tokgöz, S.B., Hamamci, I.E.: Dental enumeration and multiple treatment detection on panoramic X-rays using deep learning. Scientific Reports (2021). `https://doi.org/10.1038/s41598-021-90386-1`

31. Zhu, Q., Mathai, T.S., Mukherjee, P., Peng, Y., Summers, R.M., Lu, Z.: Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. arXiv preprint arXiv:2306.08749 (2023)