Location-Guided Automated Lesion Captioning in Whole-body PET/CT Images

Mingyang Yu[1], Yaozong Gao[2], Yiran Shu[2], Yanbo Chen[2], Jingyu Liu[2], Caiwen Jiang[1], Kaicong Sun[1], Zhiming Cui[1], Weifang Zhang[3], Yiqiang Zhan[2], Xiang Sean Zhou[2], Shaonan Zhong[4], Xinlu Wang[4], Meixin Zhao[3(✉)], Dinggang Shen[1,2,5(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai 201210, China
dgshen@shanghaitech.edu.cn
[2] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[3] Department of Nuclear Medicine, Peking University Third Hospital, Beijing 100191, China
zhaomeixin.student@sina.com
[4] Department of Nuclear Medicine, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510000, China
[5] Shanghai Clinical Research and Trial Center, Shanghai 201210, China

Abstract. Whole-body PET/CT imaging provides detailed metabolic and anatomical information, which is critical for accurate cancer staging, treatment evaluation, and radiotherapy planning. Automated lesion captioning for whole-body PET/CT is essential for reducing radiologists' workload and assisting personalized treatment decisions. Unlike previous works that focus on captioning body-part images, we propose a novel automated lesion captioning framework for whole-body PET/CT images, which usually have large volume and high anatomical variability. Our framework first leverages CLIP for lesion localization, upon which we introduce two location-guided strategies: Confidence-Guided Location Prompts (CGLP), which select top-1 or top-3 anatomical location prompts based on confidence scores to guide captioning, and Dynamic Window Setting (DWS), which applies appropriate intensity windowing to enhance visual representation of the localized regions. To our knowledge, our work is the first to achieve whole-body PET/CT lesion captioning. Experimental results on a large dataset comprising **1867** subjects from Siemens, GE, and United Imaging show that our method not only yields higher BLEU scores compared to state-of-the-art methods, but also produces consistent improvements across multiple scanner makers. This advancement has the potential to streamline radiology reporting and enhance clinical decision-making using whole-body PET/CT images.

Keywords: Lesion captioning · Whole-body PET/CT · CLIP.

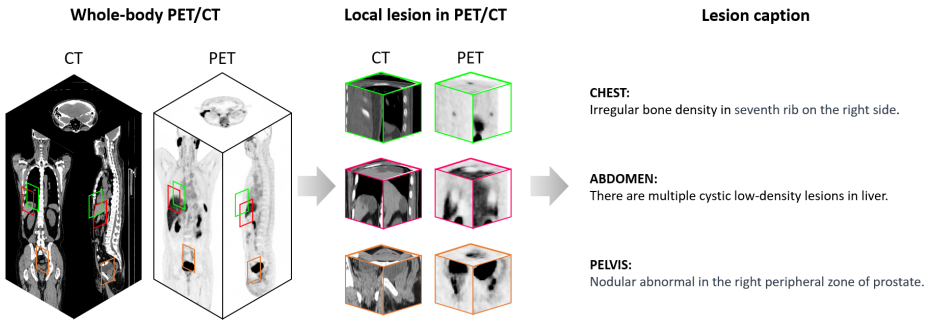

Fig. 1. Illustration of our automated lesion-captioning approach for whole-body PET/CT images. The left panel shows a 3D visualization of the full PET/CT scan, with color-coded boxes indicating suspicious lesions in the chest, abdomen, and pelvis. The middle panel shows close-up PET/CT views of these lesions, and the right panel demonstrates the automatically generated captions for each lesion.

1 Introduction

Whole-body PET/CT imaging provides high-resolution anatomical details from CT and complementary functional and molecular information from PET, making it indispensable for accurate cancer staging, treatment evaluation, and radiotherapy planning [4]. In clinical practice, generating precise lesion captions from these images is critical, *not only* for radiology report generation *but also* for assisting clinical decision-making, patient monitoring, and treatment planning [15]. However, manually captioning lesions across large anatomical regions in whole-body scans is laborious and also prone to inconsistency, underscoring the need for an automated solution.

Despite significant progress in medical image captioning, most of the existing works are designed to certain imaging modalities such as chest X-rays [17], knee X-ray [6], or breast mammography [19]. These studies typically focus on relatively small anatomical regions. In contrast, automated captioning for whole-body PET/CT images as shown in Fig. 1 remains unexplored. This gap is mainly due to two primary challenges. *First*, whole-body PET/CT images are significantly larger than those for certain regions, making it challenging to capture both global anatomical context and local lesion details simultaneously. For instance, accurately identifying and numbering similar-appearing structures such as ribs requires a large Field of View (FOV). However, since lesion caption, which includes both identification and numbering, is needed, detailed lesion information is also required, which significantly increases computational load, model complexity, and resource usage. *Second*, whole-body CT image can have large intensity dynamic range and requires contrast adjustment to account for varying tissue characteristics. The clinicians typically employ different CT window settings for tissues (such as lung and bone) to enhance the contrast between

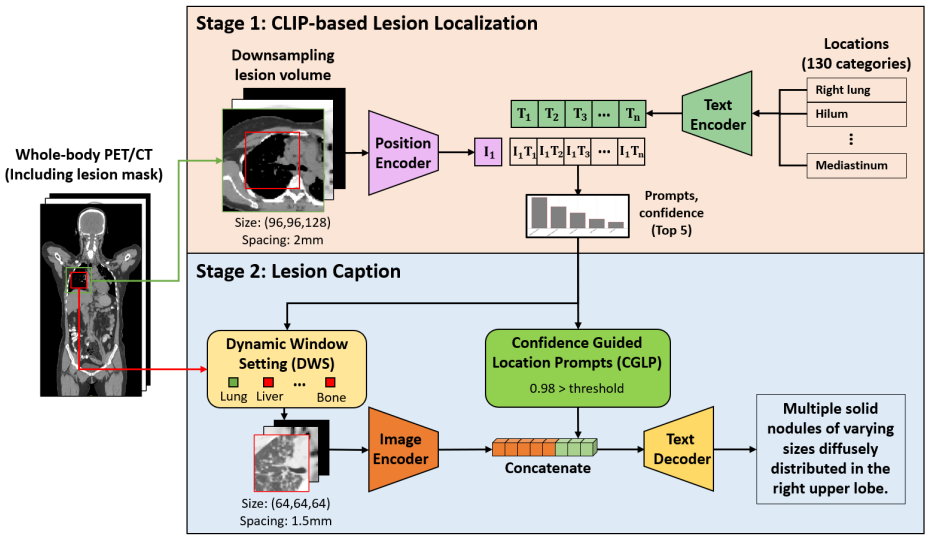

Fig. 2. Overview of our lesion captioning pipeline with two stages. In Stage 1, a CLIP-based lesion localization is conducted on a downsampled volume with a large Field of View (FOV) to retrieve the top 5 lesion locations along with their confidence values. In Stage 2, a lesion captioning process is performed, incorporating Confidence Guided Location Prompts (CGLP) for location-aware embeddings and Dynamic Window Setting (DWS) for adaptive intensity normalization to address the lack of global context in local lesion patches and enhance lesion-to-background contrast across different regions.

lesion and normal regions. An automated model that can deal with the above challenges will be highly valuable in clinical practice.

To address these challenges, we propose two strategies: 1) Confidence Guided Location Prompts (CGLP) and 2) Dynamic Window Setting (DWS). To the best of our knowledge, our work is the first to tackle automated lesion captioning for whole-body PET/CT images. Specifically, a CLIP-based lesion localization module is proposed which retrieves the top 5 potential locations along with their confidence values on the downsampled lesion volumes with large Field of View (FOV). Then, the proposed CGLP generates prompt embeddings using the potential locations and their confidence values. Meanwhile, the DWS module performs intensity normalization with the proper window setting on the CT images. These strategies enable more accurate lesion localization using global context and provide more efficient feature extraction by automatic intensity adaptation, hence enhancing the performance of lesion captioning. Extensive experiments on **5,159** lesions from multi-vendor including Siemens, GE, and United Imaging demonstrate that our approach outperforms the state-of-the-art methods by 1.2% in BLEU-4 ($p<0.05$, paired t-test), 1.9% in localization accuracy and 1.8% in CT finding.

2 Method

Our lesion captioning pipeline, as shown in Fig. 2, comprises two stages. The first stage localizes the top 5 lesion regions based on CLIP from a relative large FOV. The second stage performs lesion captioning based on the proposed Dynamic Window Setting (DWS) and Confidence Guided Location Prompts (CGLP) modules on the smaller FOV of the localized lesions. Specifically, in the first stage, the input 3D patch is aligned with 130 potential categories by CLIP, where the categories with the top 5 highest confidence are selected. In the second stage, these top 5 categories with their confidence values are further processed by the CGLP and DWS modules in an encoder–decoder lesion captioning model, where a 3D CNN-based encoder extracts volumetric features from the lesion regions and a Transformer-based decoder produces concise lesion captions. More detailed descriptions are given below.

2.1 CLIP-based Module

In our approach, the CLIP-based module [14] is employed to align the volume patch with the corresponding anatomical regions (e.g., "right lung" or "hilum"). This alignment enables the retrieval of high-confidence prompts describing the lesion locations. Specifically, for a batch of N paired samples, we use $\{\mathbf{v}_i\}_{i=1}^{N}$ to denote the image embeddings and $\{\mathbf{t}_i\}_{i=1}^{N}$ to denote the corresponding text embeddings, where $\mathbf{v}_i, \mathbf{t}_i \in \mathbb{R}^d$ and d is the embedding dimension.

The similarity between an image embedding $\mathbf{v}_i$ and a text embedding $\mathbf{t}_j$ is calculated using cosine similarity:

$$\mathrm{sim}(\mathbf{v}_i, \mathbf{t}_j) = \frac{\mathbf{v}_i^\top \mathbf{t}_j}{\|\mathbf{v}_i\| \|\mathbf{t}_j\|}. \tag{1}$$

To train the CLIP module, we adopt a symmetric contrastive loss to encourage matching of image–text pairs and penalize non-matching pairs. The loss function is defined as follows:

$$\mathcal{L}_{\mathrm{CLIP}} = \frac{1}{2}\left(\mathcal{L}_I + \mathcal{L}_T\right), \tag{2}$$

where the image-to-text loss $\mathcal{L}_I$ and the text-to-image loss $\mathcal{L}_T$ are given by

$$\mathcal{L}_I = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\mathrm{sim}(\mathbf{v}_i, \mathbf{t}_i)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathrm{sim}(\mathbf{v}_i, \mathbf{t}_j)}{\tau}\right)}, \tag{3}$$

$$\mathcal{L}_T = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\frac{\mathrm{sim}(\mathbf{t}_i, \mathbf{v}_i)}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{\mathrm{sim}(\mathbf{t}_i, \mathbf{v}_j)}{\tau}\right)}. \tag{4}$$

Here, τ is a temperature hyperparameter used to control the smoothness of the probability distribution, and $\exp(\cdot)$ denotes the exponential function. The learned CLIP-based representations form the support for our proposed CGLP and DWS in Stage 2.

2.2 Confidence Guided Location Prompts

Lesion captioning requires identifying both the lesion's location and its radiological features. It becomes especially challenging when the lesions are close to similar skeletal structures like vertebrae and ribs. For example, accurately captioning lesions near vertebrae demands the exact vertebral number, which relies on a broader anatomical view. However, although involving broader view can improve accuracy, it also increases computational costs.

To address this issue, our Confidence Guided Location Prompts (CGLP) module retrieves the top 5 anatomical locations with their corresponding confidence scores for each patch as shown in Fig. 2. Based on these confidence scores, we apply a thresholding strategy: if the highest confidence exceeds 95%, only the top 1 location is embedded; otherwise, the top 1 to top 3 location(s) are embedded. To handle varying input lengths, we use a special <pad> token for padding. The tokenized and word-embedded location information is then concatenated with the image feature, enriching it with location context and leading to more accurate lesion descriptions.

2.3 Dynamic Window Setting

CT window setting plays a crucial role in clinical practice, as different anatomical regions require distinct window parameters for better visual contrast. For example, lung regions are best visualized using a lung window with a window level of -600 and a window width of 1500, while bone structures are typically observed using a bone window (level: 800, width: 2600). For soft tissues, a soft tissue window (level: 40, width: 350) provides optimal contrast. These tailored settings enhance the visibility of specific anatomical features and improve diagnostic accuracy. Our DWS module adjusts the window settings based on the region of localized lesion in Stage 1. Specifically, when the retrieved location achieves a confidence score of 0.95 or higher, the window setting of this retrieved region is applied to the volume patch. If the confidence score falls below 0.95, the window setting of soft tissues is used. This dynamic contrast adjustment can ease the decoder and thereby thus facilitate more precise lesion description.

3 Experiments and Results

3.1 Dataset

Our dataset comprises **1,867** whole-body PET/CT scans acquired from scanners by Siemens (301), GE (290), and United Imaging (1,276), using two radiotracers: 18F-FDG [13] (1,292 cases) and 18F-PSMA [5] (575 cases). The dataset includes patients with lymphoma, nasopharyngeal carcinoma, lung cancer, prostate cancer, and liver cancer. To prevent data leakage, data are split by unique patient ID into 1,647 training cases (**46,434** lesions), 100 validation cases (4,482 lesions), and 120 test cases (**5,159** lesions). PET and CT images are normalized

Table 1. Evaluation results of our proposed model and other existing state-of-the-art models (Percentage, except CIDEr). Statistical significance test performed, with * indicating significant improvement compared to the second-best method.

Method	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
LSTM [7]	73.8	70.1	81.7	85.0	7.31	85.4
SCST [16]	77.0	73.3	83.9	86.7	7.60	87.1
AoANet [8]	77.3	73.4	84.5	87.2	7.62	87.6
LSTM-A [20]	77.6	74.1	84.5	87.3	7.67	87.7
Transformer [12]	78.1	74.6	84.5	87.2	7.71	87.6
X-LAN [10]	78.4	74.8	84.8	87.4	7.76	87.7
UpDown [2]	78.6	75.1	85.0	87.7	7.76	88.1
X-Transformer [10]	78.9	75.7	84.9	87.5	7.80	87.9
Ours	**80.1***	**76.9***	**85.8***	**88.3***	**7.92***	**88.7***

to [0,1], and standard augmentations (e.g., random rotation and scaling) are applied during training.

Lesions are segmented using an nn-UNet, followed by radiologists' review to remove false positives and ensure accurate image-text alignment. When multiple lesion types occur in the same region, each type is processed separately during training, with one lesion mask and its corresponding description per sample. The training text descriptions are derived from structured reports, which were automatically extracted from original radiology reports using a large language model (LLM), and subsequently refined by radiologists to ensure clinical accuracy and consistency with the image content.

3.2 Training Details

Training is conducted on an Nvidia L40 GPU with 48 GB memory. First, the CLIP module is trained for 300 epochs using a batch size of 48. Once it converges, its weights are frozen and the lesion caption model is trained for 300 epochs with a batch size of 64. The initial learning rate is set as 1×10^{-4}, and it is reduced by a factor of 10 every 100 epochs. The Adam is used as optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 5×10^{-4}.

3.3 Comparison with State-of-the-Art Methods

In our comparison study, we select several image captioning models whose image encoders are based on CNN architectures. The comparison includes both traditional CNN-LSTM models [7] and CNN-Transformer models [12]. Specifically, the compared methods are SCST (using LSTM) [16], LSTM-A [20], UpDown [2], AoANet [8], X-LAN [10], and X-Transformer [10]. To quantitatively evaluate the performance, we employ multiple metrics including BLEU [11], METEOR [3], ROUGE-L [9], CIDEr [18], and SPICE [1]. All the scores, except CIDEr, which remains as a raw score, are reported in percentage (e.g., 80.1% represents a score of 0.801). In Table 1, the star (*) marks indicate statistically significant

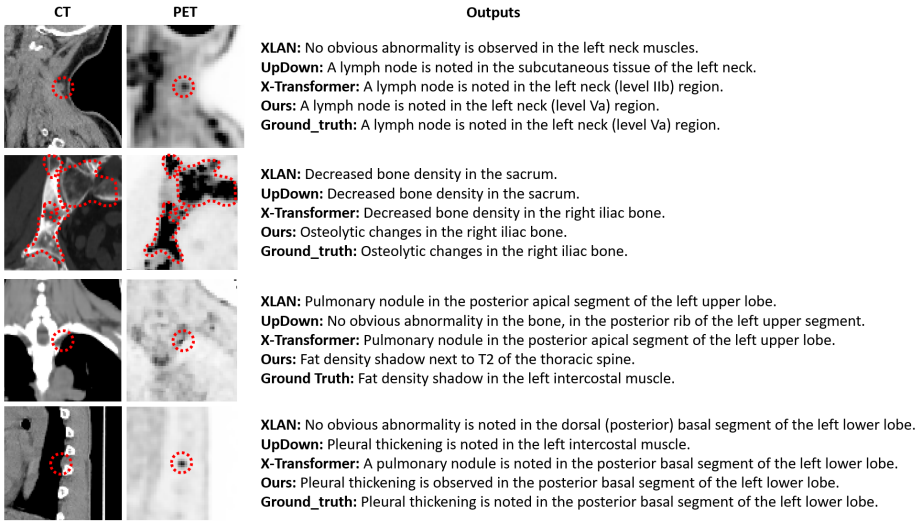

Fig. 3. Demonstration of different captioning models (Ours, XLAN, UpDown, and X-Transformer) on four representative PET/CT images. The red dotted curves in each image highlight the lesion locations, and the texts on the right show the generated descriptions by the models.

improvements of our method over the second-best performing method. A paired t-test is performed, and the p-values are all well below 0.05, confirming that our method's improvements are statistically significant across all metrics.

We summarize the performance of all the models in Table 1. We can see that the main network structures are based on LSTM and Transformer. The LSTM-based models provide robust performance. The Transformer-based models that utilize attention mechanisms consistently outperform the LSTM counterparts. Besides, we can see that X-Transformer achieves the second highest scores, indicating that the architecture of X-Transformer (X-Linear attention) significantly improves the model's ability to generate coherent and semantically rich descriptions. Our model benefits from the same model architecture as the X-Transformer, and achieves the best performance in all the evaluation metrics, i.e., in terms of BLEU, METEOR, ROUGE-L, CIDEr, and SPICE, providing the most accurate and fluent lesion descriptions among all the investigated methods.

In Fig. 3, we compare our method with three captioning models (XLAN, UpDown, and X-Transformer) using four representative PET/CT slices (left) and their generated texts (right). In the first example, our model accurately localizes the lesion in the left neck region ("level Va") thanks to the proposed CGLP module, whereas the other models misidentify the location. In the second example, our DWS module enhances subtle details, enabling correct identification of osteolytic changes rather than decreased bone density. In the third example, even though our location description slightly differs from the ground truth, it still refers to the same anatomical area. Finally, for a lesion in the left lower

Table 2. Ablation study results in percentage. To highlight clinically significant differences, we include two accuracy measures alongside traditional captioning metrics (BLEU-1 to BLEU-4).

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Localization	CT finding
B	86.3	82.3	78.9	75.7	81.0	65.5
B+W	86.5	82.6	79.3	76.0	81.2	67.3
B+P	86.8	83.0	79.7	76.5	82.2	66.2
B+W+P	**87.1**	**83.3**	**80.1**	**76.9**	**82.9**	**67.3**

lobe, our method correctly identifies both the pathology and its location, while other models only partially work. These results demonstrate the robustness of our model in both lesion localization and radiological description.

3.4 Ablation Studies

In the ablation study, we evaluate the impact of each component in our model. The baseline (B) is the standard X-Transformer, and P denotes the use of CGLP and W denotes the use of DWS. We report traditional captioning metrics (BLEU-1 to BLEU-4 in percentage) along with two accuracy measures that compute how well the model extracts location- and CT-finding-related words from the generated descriptions using word matching.

Table 2 shows that, compared to the baseline model (B), adding CGLP (B+P) slightly improves BLEU scores and notably boosts location accuracy, indicating that the prompt aids the model in better understanding lesion locations. Similarly, incorporating DWS (B+W) enhances BLEU scores and CT finding accuracy. Notably, the combined configuration (B+W+P) achieves the best performance across all the metrics, verifying the effectiveness of these two strategies to produce more accurate and context-rich captions.

Overall, the ablation study confirms that both CGLP and DWS improve the X-Transformer model, enhancing general caption quality and the targeted accuracy for location and CT finding. This evidence supports our final approach, which, built on the robust X-Transformer architecture with both CGLP and DWS, offers superior performance in generating precise and informative lesion captions.

4 Conclusion

In this paper, we present the first framework for automated lesion captioning in whole-body PET/CT images. Our method introduces two novel strategies: 1) Confidence Guided Location Prompts (CGLP), which generate coarse but informative lesion location based on confidence scores from CLIP, and 2) the Dynamic Window Setting (DWS), which dynamically adjusts CT window parameters for optimal lesion visibility. Experiments on a large dataset of **1,867**

subjects demonstrate that our model outperforms the existing methods significantly in multiple metrics including BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. Additionally, we evaluated accuracy metrics showing clinically significant differences, further confirming the advancements of our proposed model. This work marks a significant step toward automated radiological reporting for whole-body PET/CT, and can largely reduce radiologists' workload, standardize lesion annotation, and enhance clinical decision-making.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (grant numbers 82441023, U23A20295, 62131015, 82394432), the China Ministry of Science and Technology (S20240085, STI2030-Major Projects-2022ZD0209000, STI2030-Major Projects-2022ZD0213100), Shanghai Municipal Central Guided Local Science and Technology Development Fund (No. YDZX20233100001001), The Key R&D Program of Guangdong Province, China (grant number 2023B0303040001), HPC Platform of ShanghaiTech University, the special fund of Beijing Clinical Key Specialty Construction Program, P. R. China (2022), China Medical Health Development Foundation, Peking University Third Hospital and United-Imaging Research Institution Intelligential Imaging Joint Research & Development Center Foundation, and Key Clinical Project of Peking University Third Hospital (BYSYZD2019038 and BYSYZD2023016). We also acknowledge the Department of Nuclear Medicine, The First Affiliated Hospital of Guangzhou Medical University, for providing validation data.

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

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 382–398. Springer (2016)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
3. 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)
4. Blodgett, T.M., Meltzer, C.C., Townsend, D.W.: PET/CT: form and function. Radiology **242**(2), 360–385 (2007)
5. Chang, S.S.: Overview of prostate-specific membrane antigen. Reviews in urology **6**(Suppl 10), S13 (2004)
6. Gasimova, A., Montana, G., Rueckert, D.: Automated Knee X-ray Report Generation. arXiv preprint arXiv:2105.10702 (2021)
7. Hochreiter, S.: Long Short-term Memory. Neural Computation MIT-Press (1997)

8. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4634–4643 (2019)
9. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
10. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10971–10980 (2020)
11. 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)
12. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International conference on machine learning. pp. 4055–4064. PMLR (2018)
13. Phelps, M.E., Huang, S., Hoffman, E., Selin, C., Sokoloff, L., Kuhl, D.: Tomographic measurement of local cerebral glucose metabolic rate in humans with (F-18) 2-fluoro-2-deoxy-D-glucose: validation of method. Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society **6**(5), 371–388 (1979)
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. Reale-Nosei, G., Amador-Domínguez, E., Serrano, E.: From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. Medical Image Analysis p. 103264 (2024)
16. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7008–7024 (2017)
17. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest X-ray classifiers. In: BIOCOMPUTING 2021: proceedings of the Pacific symposium. pp. 232–243. World Scientific (2020)
18. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
19. Yalunin, A., Sokolova, E., Burenko, I., Ponomarchuk, A., Puchkova, O., Umerenkov, D.: Generating Mammography Reports from Multi-view Mammograms with BERT. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 153–162 (2021)
20. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Proceedings of the IEEE international conference on computer vision. pp. 4894–4902 (2017)