SGRNet: Spatially Guided Radiology Network for Structured Radiological Reporting of Head and Neck Cancer

Ayush Gupta[1,2] AYUSH.GUPTA2@ETU.UNISTRA.FR
[1] *University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France*
[2] *IHU Strasbourg, Strasbourg, France*

Vinkle Srivastav[1,2]
Prateek Upadhya[1,2]
Amit Gupta[4]
[4] *All India Institute of Medical Sciences, New Delhi, India*

Krithika Rangarajan[4]
Nicolas Padoy[1,2]

Editors: Under Review for MIDL 2026

Abstract

Automated radiological report generation has the potential to reduce reporting time and inter-observer variability. In this work, we propose SGRNet (Spatially Guided Radiology Network), a framework for generating structured radiology reports from contrast-enhanced CT (CECT) images of head and neck cancers. Using a clinically informed template designed to capture the anatomical complexity of this region, we reformulate free-text report generation as a multi-label classification problem, where the input is a CECT scan and the output is a binary label for each organ and sub-organ, indicating whether the tumor is involved. To enable effective tumor localization, we incorporate two complementary spatial priors: (1) automated organ segmentation and (2) weakly supervised tumor localization via Gaussian heatmaps. The integration of these priors substantially improves the prediction of tumor involvement, particularly in small and anatomically complex structures. We evaluate our method on a newly curated dataset of 184 paired CECT scans and corresponding reports, demonstrating that spatially guided learning significantly enhances performance. Our approach achieves a mean Average Precision (mAP) of 0.60, representing an 8.8% relative improvement over the strongest CECT-only baseline. These results highlight the potential of AI-assisted structured reporting to enable faster, more consistent, and clinically actionable assessment of head and neck cancer. The code and dataset will be made publicly available.

Keywords: Head and neck cancer, contrast-enhanced CT, structured radiology reports, weak supervision, Gaussian heatmaps, organ segmentation, 3D deep learning

1. Introduction

Advances in medical imaging have revolutionized oncology, enabling earlier detection, more precise staging, and enhanced treatment planning. However, the full clinical value of these images depends on radiological reports, which translate complex visual information into structured clinical knowledge. Despite improvements in imaging technology, radiological reporting remains predominantly manual, requiring substantial time and cognitive effort from radiologists. This gap between imaging capability and reporting capacity creates a critical bottleneck in clinical workflows, motivating the development of automated solutions for radiology report generation.

This challenge is particularly pronounced in head and neck cancer, one of the most anatomically complex disease sites in oncology. Head and neck cancer represents a significant global health burden, with 890,000 new cases and 450,000 deaths annually ((IARC), 2022). Accurate assessment requires careful evaluation of multiple anatomical subsites, including mucosal surfaces, deep fascial spaces, and regional lymph nodes, all of which are tightly packed. Radiological examinations of head and neck cancer are among the most detailed, often taking an average of 24 minutes per scan (MacDonald et al., 2013), imposing a substantial workload on clinicians. These challenges make head and neck cancer an ideal target for automated report generation.

Multiple imaging modalities support the diagnosis and treatment planning of head and neck cancer. PET/CT and hybrid MRI/PET scans provide complementary metabolic and soft-tissue information, enabling the reliable localization of tumors. However, these modalities are costly, time-intensive, and not universally available. In contrast, contrast-enhanced computed tomography (CECT) is widely accessible and routinely acquired in clinical practice. Yet, because CECT lacks direct metabolic information, tumour localisation depends on radiologists manually identifying subtle tissue changes, a process prone to inter-observer variability. Appendix A illustrates the differences between PET/CT, MRI/PET, and CECT, highlighting how PET guidance enables precise localisation, whereas tumour boundaries in CECT are visually subtle. This motivates the development of automated methods capable of accurately interpreting CECT scans.

Automated radiology report generation has seen significant progress in anatomically simpler domains. Most prior work focuses on 2D chest X-rays, leveraging large public datasets such as CheXpert (Irvin et al., 2019) to generate reports. For 3D CT scans, datasets such as CT-RATE (Hamamci et al., 2024b) exist, but they primarily cover less complex anatomical regions. Models trained on these datasets often fail to generalize to head and neck cancer due to small organ sizes, complex spatial relationships, and subtle tissue contrasts. Public datasets for head and neck cancer are limited. Available collections primarily consist of PET/CT or PET/MRI scans, which provide rich anatomical and metabolic information but often lack corresponding radiological reports, essential for supervised learning. Some datasets include tumour segmentations, but voxel-level annotations alone are insufficient for learning clinically meaningful textual descriptions. Obtaining accurate segmentations is labor-intensive and costly, requiring expert radiologists and limiting the scalability of the dataset.

To address the lack of publicly available CECT head and neck datasets paired with radiological reports, we curated a dataset of 184 head and neck cancer CECT scans, each accompanied by reports authored by a radiologist. To make these reports suitable for automated learning, we converted the free-text reports into a structured format using a clinically validated schema (Gupta et al., 2025). This conversion ensures consistency across reports, enhances clinical interpretability, and produces anatomically grounded labels capturing organ- and sub-organ tumor involvement. Given that our dataset is still relatively small to support full free-text report generation, this approach also simplifies the problem by framing it as a multi-label classification task, making it more tractable for automated learning. Learning accurate tumour representations from CECT remains challenging due to low contrast, variable tumour size, and complex anatomy. To address these challenges, we propose SGRNet, a spatially guided learning framework that incorporates two complementary priors: automated organ-level segmentation masks and weakly supervised tumor localization maps modeled as Gaussian heatmaps. Gaussian heatmaps are used because obtaining voxel-level segmentations is labor-intensive and difficult to scale. These priors are integrated through a spatial feature modulation mechanism, guiding the model to focus on anatomically relevant regions during multi-label prediction. Finally, the predicted multi-label outputs are mapped back into the structured schema to generate complete structured reports. Using our approach, we achieve an 8.8% improvement in mean average precision compared to the best baseline model, demonstrating the effectiveness of our method in handling the visual complexity of CECT scans and the scarcity of structured annotated data.

2. Related Work

Previous works in the area of radiological report generation have primarily focused on 2D imaging, particularly chest X-rays, due to the availability of large paired image–text datasets such as MIMIC-CXR (Johnson et al., 2019) (370,000+ radiographs), CheXpert (Irvin et al., 2019) (224,000+ studies), and PadChest (Bustos et al., 2020) (160,000+ annotated images). These datasets have enabled increasingly sophisticated image–text models, leading to both free-text generation (Yuan et al., 2019) and structured prediction (Wu et al., 2024). Despite their success, these approaches are limited to 2D data and do not directly extend to volumetric 3D modalities such as CT or MRI, which are essential for capturing the complex, 3D anatomical structures inherent to head and neck cancer.

Recent studies have begun to tackle automatic report generation for 3D imaging, driven largely by the release of datasets like CT-RATE (Hamamci et al., 2024b), which contains 25,692 non-contrast CT scans paired with free-text reports, predominantly of thoracic anatomy. Leveraging this dataset, (Hamamci et al., 2024a) introduced CT2Rep, using a spatial encoder and transformer-based decoder, while (Chen et al., 2024) proposed a ViT-based 3D encoder coupled with a transformer decoder. Although these methods demonstrate the feasibility of 3D report generation, they focus mainly on chest CT and do not address the anatomical complexity or reporting requirements of head and neck imaging.

For head and neck cancer specifically, publicly available datasets are more limited. Most existing resources, particularly those on TCIA (Clark, Kenneth and Vendt, 2013), are based on PET/CT. While PET/CT is effective for tumor localization, it requires radioactive tracers, making the procedure costly and time-consuming. For instance, a recent multicenter dataset by (Saeed and Hassan, 2025) comprises 1,123 PET/CT studies with expert segmentations of primary tumors (GTVp) and involved lymph nodes (GTVn), along with clinical metadata such as TNM staging and HPV status. While these datasets are invaluable for segmentation and outcome prediction, they do not include contrast-enhanced CT (CECT) imaging paired with radiology reports, thereby limiting research on automated report generation from the more accessible CECT modality. Similarly, while (Walter et al., 2024) trained a nn-UNet model to segment 71 H&N structures, the resulting segmentations serve only as anatomical priors in our work, and the study (Walter et al., 2024) focuses purely on segmentation rather than the multi-label structured report generation task.

Weak supervision has emerged as a promising approach to alleviate the annotation burden in medical imaging. Techniques using point annotations, Gaussian blobs, and heatmaps have improved segmentation performance with minimal manual effort. For instance, (Roth et al., 2021) converted extreme points on object boundaries into Gaussian blobs to guide 3D medical segmentation, (Qu et al., 2020) expanded sparse point annotations in histopathology into Gaussian masks, and (Zhong et al., 2024) used radiologists' gaze points to generate heatmaps as spatial priors. To our knowledge, the integration of Gaussian heatmaps as a spatial prior for tumor localization has not been applied to the automatic generation of structured radiological reports, representing an opportunity to improve CECT-based head and neck workflows.

3. Dataset

We present a multicentric dataset of 184 contrast-enhanced CT (CECT) scans of patients with head and neck cancer. This cohort was curated from two sources: 49 scans from the publicly available ACRIN-HNSCC study (Kinahan et al., 2019) and 135 scans collected at the All India Institute of Medical Sciences (AIIMS), New Delhi. Data curation involved filtering the original 260 ACRIN-HNSCC patients to select usable CECT series and assess scan quality. All data samples were fully de-identified prior to use to ensure patient privacy and compliance with ethical standards. The dataset exhibits substantial heterogeneity in acquisition protocols, reflecting real-world clinical variability across the two centers. Tumors span common head and neck sub-sites, including the supraglottis, transglottis, nasopharynx, and oropharynx, ensuring broad clinical relevance. Scans were acquired on multiple devices, including SIEMENS SOMATOM Definition AS, GE MEDICAL SYSTEMS BrightSpeed, and Philips Brilliance 16, with tube voltages (kVp) ranging from 100 to 120. Slice thickness varied between 0.5mm and 1.5mm. We summarize the core imaging parameters in Table 3.

The annotation process was two-fold to achieve a fully structured dataset. Scans from the AIIMS dataset (N = 135) were initially paired with their corresponding original free-text radiology reports (average length, 275 words). For the ACRIN-HNSCC scans (N = 49), which lacked corresponding reports, our expert radiologist generated new structured reports using the standardized template to ensure consistency across the entire dataset. The AIIMS free-text reports were subsequently converted into this same structured format via the procedure detailed in (Gupta et al., 2025). Furthermore, a crucial subset of the AIIMS cohort (N = 39) was separately annotated with weak bounding-box cuboids centered on the primary tumor; these were used exclusively to train our weak tumor localization prior (Section 4.3.2) and were excluded from the main classification training to prevent data leakage. The resulting core modeling dataset, comprising 145 scans (184 total, minus the 39 scans reserved for weak supervision), was used for all training and validation purposes. It was split via stratified five-fold cross-validation to ensure balanced representation of tumor sub-sites.

4. Method

To address the problem of radiological report generation for head and neck cancers, we first reformulate the task as a structured report generation problem. Rather than producing lengthy free-text descriptions, we utilize a standardized template that organizes findings into predefined fields corresponding to clinically relevant organs and their respective sub-organs. This reformulation converts the task into a multi-label classification problem, where the objective is to predict, for each anatomical structure, whether it is involved or affected by the tumor. We begin by establishing baseline models for this structured prediction task using only the raw CECT volumes as input. We then seek to improve upon these baselines by incorporating spatial priors derived from anatomical and tumor localization. Specifically, we propose SGRNet (detailed in Section 4.3) that integrates these two priors, organ segmentation maps, and weakly supervised Gaussian tumor heatmaps, allowing the model to focus explicitly on clinically relevant anatomical regions and achieve substantial improvements in prediction.

4.1. Conversion of Free-Text Radiological Reports to Structured Reports

Free-text radiological reports are converted into structured labels using the prompting strategy of (Gupta et al., 2025), where GPT-4 is guided to populate a predefined hierarchical template. A two-step prompting process is used to generate and refine reports, with unsupported fields marked as *"Missing"*. All structured reports are reviewed by expert radiologists and used as ground-truth labels. This process reformulates the task as a multi-label classification problem over anatomical structures.

4.2. Baselines

We evaluate two standard 3D convolutional baselines: DenseNet121 (Huang et al., 2017) and EfficientNet-B0 (Tan and Le, 2019), implemented in MONAI (Cardoso et al., 2022). Both models operate on preprocessed CECT volumes and use global average pooling followed by a sigmoid-based multi-label classification head.

We also evaluate CT-FM (Pai et al., 2025), a foundation model that generates global CT volume embeddings, which are then adapted using a lightweight multilayer perceptron and a sigmoid classification head. All baselines use only CECT volumes as input, without organ segmentation masks or Gaussian tumour priors, enabling fair comparison with our proposed method.

4.3. Overview of SGRNet

We propose SGRNet (Spatially Guided Radiology Network), a multi-branch 3D architecture that integrates anatomical and tumor-related spatial priors to guide the generation of structured reports from contrast-enhanced CT (CECT) scans. The model takes as input the raw CECT volume, which is used to generate automated organ segmentation masks via a pre-trained nn-UNet model, as well as Gaussian heatmaps providing weak tumour localisation. Each input is encoded through a dedicated 3D encoder to produce latent feature maps. The organ and tumour features are combined to generate a spatial modulation map, acting as a soft gating mechanism on the CECT features and emphasising regions consistent with anatomical and pathological priors. The modulated CECT features are then fused with the prior feature maps and passed through a 3D convolutional fusion block followed by a multilayer perceptron to produce multi-label predictions. These predictions

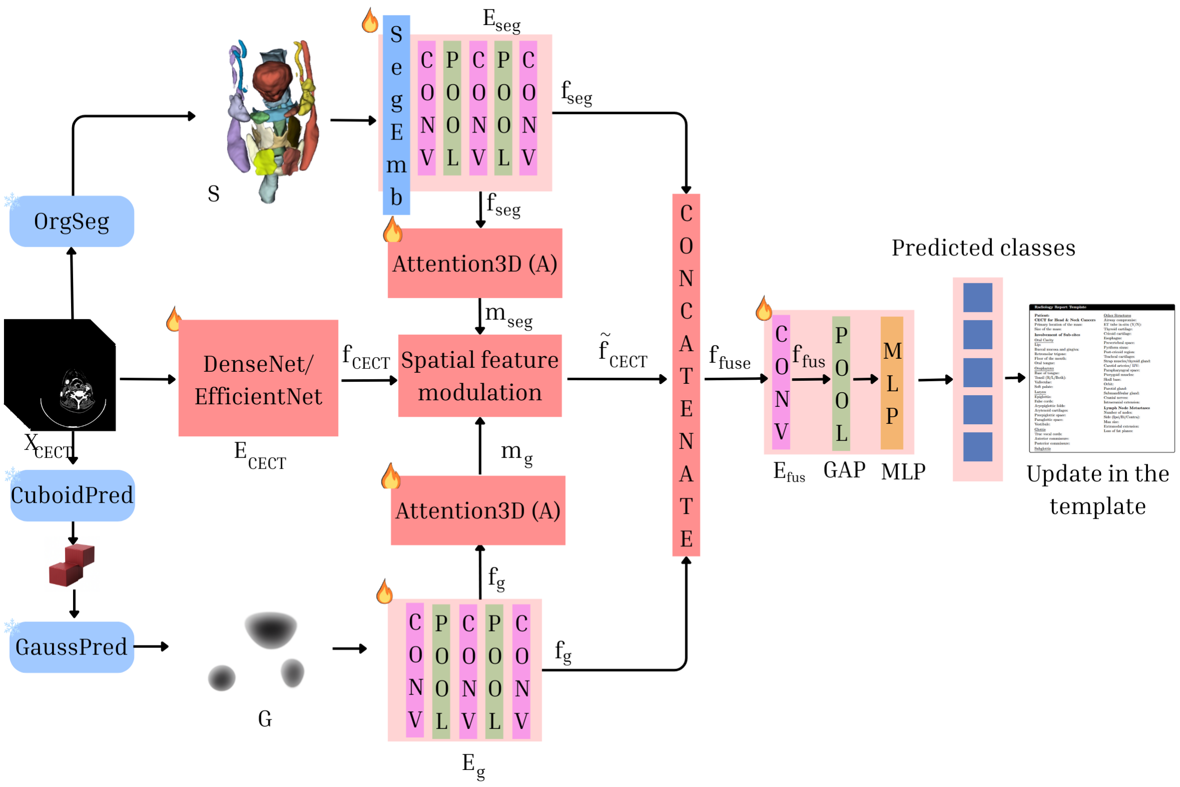

Figure 1: Architecture of the SGRNet. The tumour heatmap and the organ segmentation stream modulate the CECT stream via spatial feature modulation, allowing the network to emphasise tumour-salient regions while preserving global anatomical context.

are subsequently used to populate the structured reporting template, completing the automated structured report generation process. In the following, we describe each component in detail.

4.3.1. Automated Organ Segmentation (OrgSeg)

Organ segmentation masks are generated using TotalSegmentator, which is based on a pre-trained nn-UNet (Isensee et al., 2021) with publicly available weights from (Walter et al., 2024). From the full set of 71 anatomical structures, we select 21 regions relevant to head and neck cancer (see Appendix E). The resulting integer label map

$$S \in \{0, \ldots, 21\}^{D \times H \times W}$$

represents the segmented volume, where D, H, and W denote the depth, height, and width of the input CECT scan, respectively, and 0 denotes the background.

To avoid creating a high-dimensional one-hot encoded volume (i.e., one channel per segmentation class), we embed the single-channel label map into a low-dimensional space using an embedding function:

$$E_{emb}(S) = \text{Embed}(S) \in \mathbb{R}^{d_e \times D \times H \times W},$$

where d_e is the embedding dimension. This approach allows the network to operate on a single-channel input with values ranging from 0 to 21 while learning compact semantic representations for each anatomical region. The embedded volume is then processed by the segmentation encoder (E_{seg}) to produce the anatomical prior feature map:

$$f_{seg} = E_{seg}(E_{emb}(S)) \in \mathbb{R}^{C_{seg} \times d \times h \times w},$$

where C_{seg} is the number of output channels and d, h, w are the downsampled spatial dimensions of the encoded feature map. This feature map is subsequently used in the fusion stage to guide the model's attention to relevant anatomical structures.

4.3.2. Weak Tumour Localisation via Gaussian Heatmaps (CuboidPred+GausPred)

A nn-UNet model is pre-trained on the dataset introduced in (Saeed and Hassan, 2025) using only the CT channel of the PET/CT scans, and subsequently fine-tuned on 39 cases from our cohort for coarse tumour localisation. The model predicts binary cuboid regions, where voxels inside the tumour bounding box are assigned 1 and background voxels 0.

To convert the predicted cuboid regions into Gaussian heatmaps, let $\Omega \subset \mathbb{R}^3$ denote the set of connected foreground components after thresholding and size filtering. For each component, an axis-aligned cuboid is fitted. The centre of the cuboid, (μ_x, μ_y, μ_z), is computed as the centre of mass of the connected component, and the cuboid's side lengths $(w_{\text{cuboid}}, h_{\text{cuboid}}, d_{\text{cuboid}})$ along the x, y, and z axes are determined based on the spatial extent of the component.

A 3D Gaussian heatmap is generated as:

$$G(x, y, z) = \exp\left(- \Big(\frac{(x - \mu_x)^2}{2\sigma_x^2} + \frac{(y - \mu_y)^2}{2\sigma_y^2} + \frac{(z - \mu_z)^2}{2\sigma_z^2} \Big) \right),$$

where $\sigma_x, \sigma_y, \sigma_z$ are the standard deviations along each axis, proportional to the cuboid's width, height, and depth, $(w_{\text{cuboid}}, h_{\text{cuboid}}, d_{\text{cuboid}})$respectively. The voxel coordinates in the 3D volume are denoted by x, y, z, and the mean of the Gaussian is set to the cuboid centre (μ_x, μ_y, μ_z).

The resulting Gaussian volume G is encoded by the Gaussian encoder E_g to produce a latent feature map:

$$f_g = E_g(G) \in \mathbb{R}^{C_G \times d \times h \times w},$$

where C_G is the number of feature channels, and d, h, w are the spatial dimensions of the encoded feature map. This feature map serves as a weak spatial prior for tumour localisation and is subsequently used in the spatial feature modulation and fusion stages.

4.3.3. Spatial Feature Modulation

To integrate anatomical and tumour priors with imaging features, spatial attention maps are generated using a lightweight 3D attention block $A(\cdot)$, composed of $1\times1\times1$ convolutions followed by a sigmoid:

$$m_{seg} = A(f_{seg}), \quad m_g = A(f_g), \quad m = m_{seg} + m_g,$$

where $m_{seg}, m_g, m \in \mathbb{R}^{1\times d\times h\times w}$ and (d, h, w) denote the depth, height, and width of the encoded feature maps. The CECT volume X_{CECT} is encoded using a DenseNet/EfficientNet encoder E_{CECT}:

$$f_{\text{CECT}} = E_{\text{CECT}}(X_{\text{CECT}}) \in \mathbb{R}^{C_{\text{CECT}}\times d\times h\times w},$$

where C_{CECT} is the number of feature channels. Spatial modulation is then applied as:

$$\tilde{f}_{\text{CECT}} = f_{\text{CECT}} \odot (1 + m),$$

where $\odot$ denotes element-wise multiplication. This operation acts as a soft spatial gating mechanism that enhances anatomically and tumour-consistent regions while preserving global context.

4.3.4. Feature Fusion and Classification

The modulated CECT features $\tilde{f}_{\text{CECT}}$ are concatenated with the anatomical and tumour prior feature maps f_{seg} and f_g to form a unified representation, which is passed through a 3D convolutional fusion block E_{fus}, pooled, and fed into an MLP with sigmoid activation to produce multi-label predictions:

$$\hat{y} = \sigma\Big(\text{MLP}\big(\text{GAP}(E_{fus}(\text{Concat}(\tilde{f}_{\text{CECT}}, f_{seg}, f_g)))\big)\Big),$$

where $\hat{y} \in [0, 1]^N$ represents the predicted labels for each anatomical structure in the structured reporting template, and N is the total number of labels. This final prediction is then passed to the template to complete the automated structured report generation pipeline.

5. Experiments & Results

5.1. Implementation Details

All models are trained on 145 contrast-enhanced CT (CECT) scans paired with structured radiological reports, using 5-fold cross-validation with stratified sampling. The prediction task is restricted to the five most consistently annotated structures: *Tongue*, *Hypopharynx*, *Larynx_air*, *Strap muscles + Thyroid gland*, and *Carotid arteries + Internal jugular vein.* CECT volumes are resampled to 4 mm isotropic resolution, clipped to a soft-tissue window of $[-250, 150]$ HU, and normalized per volume. The average input size is $128 \times 128 \times 171$ voxels.

Segmentation embeddings have dimension $d_e = 8$, and all encoder feature maps have spatial size $42 \times 32 \times 32$ with 64 channels. Standard 3D data augmentation is applied, and models are trained end-to-end with the Adam optimizer (learning rate 1×10^{-5}, batch size

2) using early stopping based on the validation mAP. Class imbalance is addressed via the Asymmetric Loss (ASL), defined as:

$$\mathcal{L}_{\text{ASL}} = -\sum_{c=1}^{K} \Big[y_c \, (1-p_c)^{\gamma_+} \, \log(p_c) + (1-y_c) \, p_c^{\gamma_-} \, \log(1-p_c) \Big],$$

where y_c and p_c are the ground-truth label and predicted probability for class c, K is the number of classes, and $\gamma_+ = 1, \gamma_- = 2$ in our experiments. Performance is reported using mean Average Precision (mAP).

For weak tumour localization, a nn-UNet predicts coarse 3D cuboid masks, pre-trained on a public PET/CT dataset and fine-tuned on radiologist annotations using 5-fold cross-validation. Fine-tuning is performed for 250 epochs with learning rate decay. Predicted cuboids are converted into Gaussian heatmaps as described in Section 4.3.2.

5.2. Results

Table 1 summarizes the Average Precision (AP) for each anatomical structure across model configurations. Baseline models trained solely on CECT achieved moderate performance, with mean APs of 0.512 (DenseNet), 0.494 (EfficientNet), and 0.488 (CT-FM). CT-FM performs better on larger structures but underperforms on smaller soft-tissue regions. SGRNet, by incorporating organ-based and weak tumour localisation via Gaussian heatmaps, consistently improves AP for both DenseNet and EfficientNet, particularly in complex regions such as the hypopharynx and laryngeal airway, with DenseNet achieving a mean AP of 0.600. These results indicate that combining anatomical and pathological priors enhances the prediction of structured reports in regions affected by tumor-induced distortions. A full ablation study of spatial prior strategies is included in Table 2.

For weak tumor localization, nn-UNet achieves a mean Dice score of 0.37 when trained from scratch, which improves to 0.51 with external CT pretraining. Example outputs are provided in Appendix F

Model	**Input**	**Tongue AP**	**Hypopharynx AP**	**Larynx Air AP**	**Strap + Thyroid AP**	**Carotid + IJV AP**	**Mean AP**
DenseNet	CECT	0.549	0.559	0.574	0.357	0.519	0.512
EfficientNet	CECT	0.543	0.534	0.563	0.339	0.489	0.494
CT-FM	CECT	0.419	0.576	0.612	0.256	0.579	0.488
SGRNet(DenseNet)	Org+Tum-Modul	0.478	0.710	0.758	0.501	0.554	0.600
SGRNet(EfficientNet)	Org+Tum-Modul	0.498	0.674	0.724	0.467	0.558	0.584

Table 1: Average Precision (AP) for baseline CECT models and spatial prior variants. Column abbreviations: IJV: Internal jugular vein.

6. Discussion

Incorporating anatomical and tumour-aware spatial priors substantially improves structured report prediction. Baseline CECT-only models demonstrate moderate performance, with

Model	**Input**	**Tongue AP**	**Hypo-pharynx AP**	**Larynx Air AP**	**Strap + Thyroid AP**	**Carotid + IJV AP**	**Mean AP**
				Organ Priors			
DenseNet	Org-Add	0.504	0.725	0.699	0.314	0.627	0.574
DenseNet	Org-Modul	0.494	0.674	0.756	0.447	0.601	0.594
EfficientNet	Org-Add	0.475	0.708	0.658	0.283	0.593	0.543
EfficientNet	Org-Modul	0.478	0.636	0.701	0.413	0.600	0.567
				Organ + Tumour Priors			
DenseNet	Org+Tum-Add	0.485	0.731	0.715	0.367	0.614	0.582
EfficientNet	Org+Tum-Add	0.431	0.713	0.686	0.345	0.543	0.544

Table 2: Ablation study showing the effect of different spatial prior integration strategies (Add, Modulation) for organ and tumour priors. Abbreviations: IJV – Internal jugular vein.

DenseNet and EfficientNet exhibiting similar trends, while CT-FM performs better on larger anatomical structures. Combined organ- and tumour-based spatial modulation in SGRNet consistently enhances AP, particularly in anatomically complex regions. This combination of priors enables the network to better handle tumour-induced anatomical distortions. For weak tumour localisation, nnU-Net benefits significantly from CT-based pretraining (mean Dice score of 0.51 vs. 0.37). This suggests that standard CT data can effectively transfer anatomical knowledge to CECT tasks, highlighting the potential of large-scale PET/CT datasets using only their CT component as a valuable resource for training models and opening new directions for leveraging such data in contrast-enhanced imaging tasks. Key limitations remain, including the relatively small dataset size, the coarse nature of cuboid-based tumour annotations, and potential errors in automated organ segmentations.

7. Conclusion

We present a framework for generating structured radiological reports in head and neck cancer using contrast-enhanced CT (CECT) scans. By incorporating spatial priors from organ segmentation and weak tumor localization, our model, SGRNet, improves the multi-label prediction of tumor involvement, particularly for small and anatomically complex structures. The combination of anatomical and tumor priors enables the network to focus on clinically relevant regions, yielding consistent gains over CECT-only baselines. Pretraining on external CT data further enhances segmentation and localization performance, reducing reliance on costly imaging modalities such as PET. Future work includes expanding the dataset across institutions to improve generalization, leveraging semi-supervised learning to exploit unannotated volumes, and exploring anomaly detection to identify unexpected pathological changes. Overall, structured incorporation of anatomical and pathological priors is a promising strategy for enhancing automated radiological reporting in complex anatomical regions.

8. Acknowledgements

This work was partially supported by the Interdisciplinary Thematic Institute HealthTech, as part of the ITI 2021-2028 program of the University of Strasbourg, CNRS, and Inserm, funded by IdEx Unistra (ANR-10-IDEX-0002) and SFRI (STRAT'US project, ANR-20-SFRI-0012) under the framework of the French Investments for the Future Program. This work was also partially supported by French state funds managed by the ANR under Grant ANR-10-IAHU-02. This work was granted access to the HPC resources of IDRIS under the allocations AD011011631R4 made by GENCI. The authors would like to acknowledge the High-Performance Computing Center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources was funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data.

References

Aurelia Bustos, Antonio Pertusa, Jose M Salinas, and Maria de la Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Ni Ma, Eric Kerfoot, Yipeng Wang, Andriy Myronenko, Can Zhao, Ziyue Zhang, Daguang Chen, et al. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Hao Chen, Wei Zhao, Yingli Li, Tianyang Zhong, Yisong Wang, Youlan Shang, Lei Guo, Junwei Han, Tianming Liu, Jun Liu, and Tuo Zhang. 3d-ct-gpt: Generating 3d radiology reports through integration of large vision-language models. *arXiv preprint arXiv:2409.19330*, 2024. URL https://arxiv.org/abs/2409.19330.

Clark, Kenneth and Vendt. The cancer imaging archive (tcia). https://www.cancerimagingarchive.net/, 2013.

Amit Gupta, Hema Malhotra, Amit K. Garg, and Krithika Rangarajan. Enhancing radiological reporting in head and neck cancer: Converting free-text ct scan reports to structured reports using large language models. *Indian Journal of Radiology and Imaging*, 35(1): 043–049, 2025. doi: 10.1055/s-0044-1788589.

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012, pages 476–486. Springer Nature Switzerland, 2024a. doi: 10.1007/978-3-031-72390-2_45. URL https://doi.org/10.1007/978-3-031-72390-2_45.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Doga, Omer Faruk Durugol, Weicheng Dai, Murong Xu, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024b. URL https://arxiv.org/abs/2403.17834. Dataset: CT-RATE — 25,692 non-contrast chest CT volumes (50,188 after reconstructions) from 21,304 unique patients ("non-commercial, research only" license CC BY-NC-SA).

Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens Van Der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4700–4708, 2017.

(IARC). Globocan 2022: Global cancer statistics. https://gco.iarc.fr/, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jared Seekins, David A Mong, Safwan S Halabi, Jacob K Sandberg, Russell Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, 2019.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Paul Kinahan, Michael Muzi, Brian Bialecki, and Lisa Coombs. Data from the acrin6685 trial hnscc-fdg-pet/ct. The Cancer Imaging Archive (TCIA) Collection: ACRIN-HNSCC-FDG-PET-CT, 2019. Accessed via TCIA. :contentReference[oaicite:0]index=0.

Sharyn L. S. MacDonald, Ian A. Cowan, Richard A. Floyd, and Rob Graham. Measuring and managing radiologist workload: measuring radiologist reporting times using data from a radiology information system. *Journal of Medical Imaging and Radiation Oncology*, 57(5):558–566, 2013. doi: 10.1111/1754-9485.12092.

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001*, 2025. URL https://arxiv.org/abs/2501.09001.

Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M. Riedlinger, Subhajyoti De, Shaoting Zhang, and Dimitris N. Metaxas. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11):3655–3666, 2020. doi: 10.1109/TMI.2020.3002244.

Holger R. Roth, Dong Yang, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Going to extremes: Weakly supervised medical image segmentation. *Machine Learning and Knowledge Extraction*, 3(2):507–524, 2021. doi: 10.3390/make3020026.

Numan Saeed and Hassan. A multimodal and multi-centric head and neck cancer dataset for segmentation, diagnosis and outcome prediction. *arXiv preprint arXiv:2509.00367v3*, 2025. URL https://arxiv.org/abs/2509.00367v3.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

Alexandra Walter, Philipp Hoegen-Saßmannshausen, Goran Stanic, Joao Pedro Rodrigues, Sebastian Adeberg, Oliver Jäkel, Martin Frank, and Kristina Giske. Segmentation of 71 anatomical structures necessary for the evaluation of guideline-conforming clinical target volumes in head and neck cancers. *Cancers*, 16(2):415, 2024. doi: 10.3390/cancers16020415. URL https://www.mdpi.com/2072-6694/16/2/415.

Xinyu Wu et al. Maira-2: Grounded chest x-ray report generation with radfact evaluation. *arXiv preprint arXiv:2406.04449*, 2024. URL https://arxiv.org/abs/2406.04449.

Jianbo Yuan, Hanyu Liao, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. *arXiv preprint arXiv:1907.09085*, 2019. URL https://arxiv.org/abs/1907.09085.

Yuan Zhong, Chenhui Tang, Yumeng Yang, Ruoxi Qi, Kang Zhou, Yuqi Gong, Pheng-Ann Heng, Janet H. Hsiao, and Qi Dou. Weakly-supervised medical image segmentation with gaze annotations. In *Medical Image Computing and Computer–Assisted Intervention – MICCAI 2024*, volume 15003 of *Lecture Notes in Computer Science*, pages 530–540. Springer, 2024. doi: 10.1007/978-3-031-72384-1_50.

Appendix A. Comparing PET/CT, PET/MRI and CECT

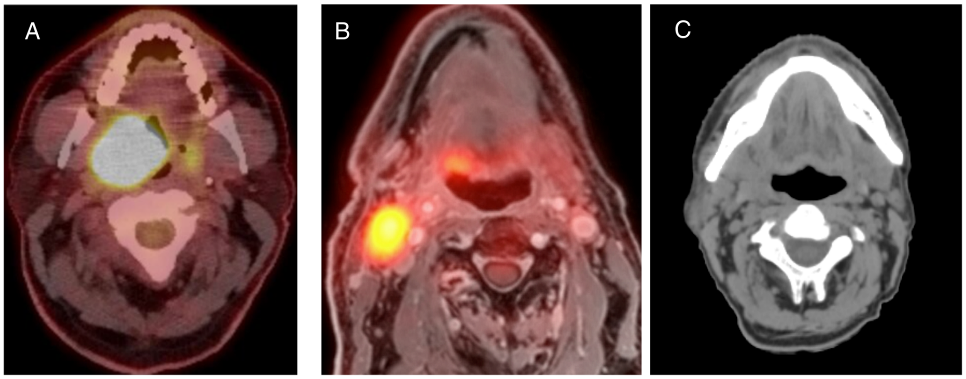

Figure 2: Visual comparison of imaging modalities. (A) PET/CT highlights metabolically active tumour regions via radiotracer uptake. (B) MRI/PET provides high soft-tissue contrast with PET-guided localisation. (C) CECT lacks metabolic tracers, resulting in more subtle tumour boundaries and motivating the need for automated interpretability methods.

Appendix B. Example of a free text report

Patient 1

Procedure: CECT of Face & Neck.

Clinical background: Ca oropharynx (cT4aN3bM0) post CT/RT

There is ill-defined enhancing soft tissue thickening in oropharynx measuring approx. 2.7cm in longest dimension involving the right lateral and posterior oropharyngeal wall, right vallecula, median glossoepiglottic fold, the tip of the epiglottis, and right base of tongue. Non-enhancing edematous thickening of bilateral aryepiglottic fold and vocal cords is noted likely post RT changes. The hyoid bone is normal. The prevertebral fat space is maintained. Enlarged multiple conglomerated necrotic lymph nodes in bilateral level II, III regions; the largest measuring ∼2x2.4cm in LAD on the right side. There is encasement of vessels on the right side with thrombosis of IJV from the level of the jugular foramen extending to the sigmoid sinus.

Paranasal sinuses/nasal cavity: Mucosal thickening is seen in the right maxillary sinus.

Maxilla/mandible: Normal.

Infra-temporal neck spaces: Normal.

Additional information: None

Comparison: Compared to the scan dated 17 June 2022, there is a reduction in the size of the mass and nodes.

Impression: Present scan shows irregular enhancing thickening involving the oropharynx with extensions as described with bilateral cervical conglomerated necrotic lymphadenopathy.

Appendix C. Structured template for head and neck cancers

Radiology Report Template

Patient:

CECT for Head & Neck Cancers

Primary location of the mass:
Size of the mass:

Involvement of Sub-sites

Oral Cavity

Lip:
Buccal mucosa and gingiva:
Retromolar trigone:
Floor of the mouth:
Oral tongue:

Oropharynx

Base of tongue:
Tonsil (R/L/Both):
Valleculae:
Soft palate:

Larynx

Epiglottis:
False cords:
Aryepiglottic folds:
Arytenoid cartilages:
Preepiglottic space:
Paraglottic space:
Vestibule:

Glottis

True vocal cords:
Anterior commissure:
Posterior commissure:

Subglottis

Other Structures

Airway compromise:
ET tube in-situ (Y/N):
Thyroid cartilage:
Cricoid cartilage:
Esophagus:
Prevertebral space:
Pyriform sinus:
Post-cricoid region:
Tracheal cartilages:
Strap muscles/thyroid gland:
Carotid arteries/ IJV:
Parapharyngeal space:
Pterygoid muscles:
Skull base:
Orbit:
Parotid gland:
Submandibular gland:
Cranial nerves:
Intracranial extension:

Lymph Node Metastases

Number of nodes:
Side (Ipsi/Bi/Contra):
Max size:
Extranodal extension:
Loss of fat planes:

Appendix D. Summary of dataset characteristics

Parameter	Value	Unit
Total Scans (N)	184	–
Average Voxel Spacing	$1 \times 1 \times 1$	mm^3
Average Resolution (X×Y)	512×510	pixels
Average Slices (Z)	288	slices
Average Mean HU	-763 (Std. 576)	HU
HU Range (Min to Max)	$[-1732, 3444]$	HU

Table 3: Summary of dataset-level imaging characteristics.

Appendix E. Details of the organs from TotalSegmentator being segmented

Label ID	Organ
0	Background
1	Larynx air
2	Thyroid cartilage
3	Cricoid cartilage
4	Hyoid bone
5	Tongue
6	Digastric (left)
7	Digastric (right)
8	Sternothyroid (left)
9	Sternothyroid (right)
10	Thyrohyoid (left)
11	Thyrohyoid (right)
12	Submandibular gland (left)
13	Submandibular gland (right)
14	Thyroid gland
15	Internal carotid artery (left)
16	Internal carotid artery (right)
17	Internal jugular vein (left)
18	Internal jugular vein (right)
19	Trachea
20	Oropharynx
21	Hypopharynx

Table 4: List of segmented organs and their corresponding label IDs from the TotalSegmentator dataset.

Appendix F. Outputs of the Anatomical Organ Segmentation and Tumour Segmentation Models

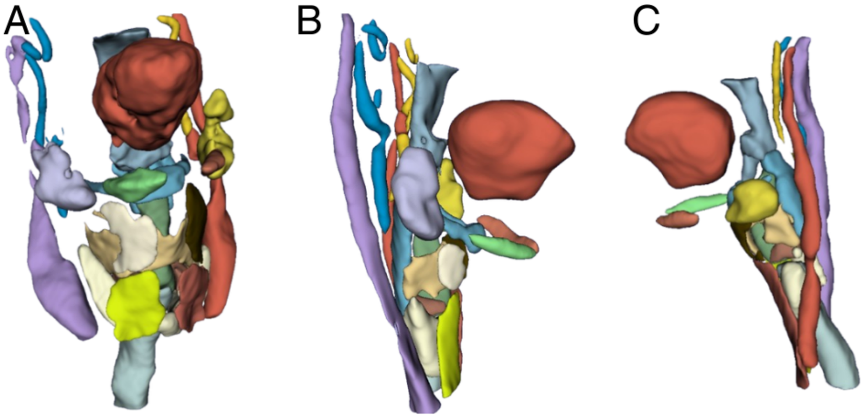

Figure 3: Qualitative results from the anatomical organ segmentation model. Panels (A), (B), and (C) show example outputs from three different patients, illustrating the predicted masks for the 21 anatomical regions defined in our pipeline.

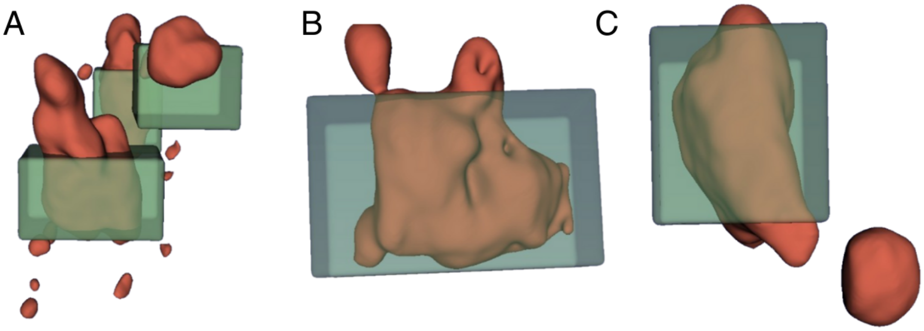

Figure 4: Tumour segmentation results from an nnU-Net model trained solely on CT images from the public PET/CT dataset. Panels (A), (B), and (C) display outputs for three representative patients.

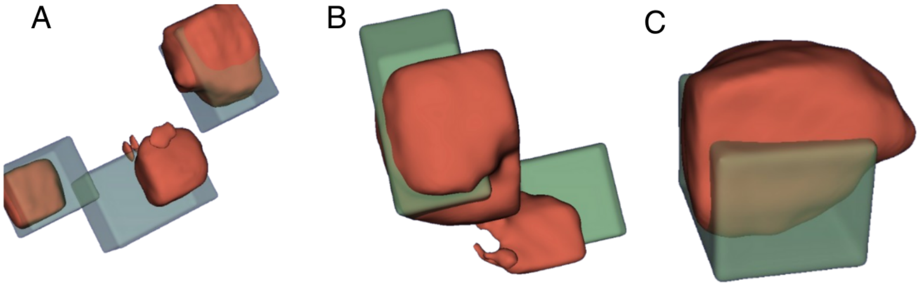

Figure 5: Tumor segmentation results from the nnU-Net model, pre-trained on CT scans from the public PET/CT dataset and subsequently fine-tuned on our curated cuboid regions. Panels (A), (B), and (C) show outputs for three different patients.

Appendix G. Equations

Dice Score. The Dice similarity coefficient between two binary masks A and B is defined as:

$$\mathrm{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|},$$

where A and B denote the predicted and ground-truth cuboid masks, respectively.