Towards Explainable Segmentation of Complex Boundaries in Lung Nodule Detection

Mohammad Asifur Rahim
Information and Computer Science Department
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
g202417400@kfupm.edu.sa

Mufti Mahmud
Information and Computer Science Department
SDAIA-KFUPM Joint Research Center for AI
Interdisciplinary Research Center for Bio Systems and Machines
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
mufti.mahmud@kfupm.edu.sa, muftimahmud@gmail.com

Abstract

Many deep learning models are computationally expensive while capturing complex edges in tasks such as lung nodule segmentation from 2D CT scans. Also, the lack of explainability hinders their adoption for clinical use. To address these challenges, this work proposes a Sobel-enhanced edge-aware powered U-Net architecture capable of emphasising the edges of nodules in lung computed tomography images. The model is trained and evaluated on the benchmark LIDC-IDRI dataset. To provide interpretability, four post hoc explainers were employed: Grad-CAM, Score-CAM, Layer-CAM, and Counterfactual explainability. The proposed model achieved competitive performance across several metrics, including accuracy, dice score, intersection over Union, sensitivity, and specificity, when compared with three baseline models– U-Net, ResUnet++, and U-Net++. Although it has slightly more parameters (3.4 million) than the U-Net (3.3 million), its ability to identify complex edges of lung nodules makes it stand out. Moreover, the four explainers effectively generated heatmaps that highlight the detected edges. The proposed model delivers competitive segmentation performance with improved edge detection and explainability, highlighting its potential for clinical deployment.

1 Introduction

Lung cancer is the second most common cancer worldwide, with more than 230,000 new cases annually (10). It is one of the major causes of cancer deaths worldwide, making early detection of lung nodules very important. Lung nodules are small growths in the lungs that can be harmless or cancerous. Computed Tomography (CT) scans can spot these nodules, but accurate segmentation—measuring their size, shape, and position—is key. This helps doctors tell if a nodule is benign or malignant and plan the right treatment, especially for cancer cases where staging and therapy choice depend on precise segmentation (1). Several AI developments have been done to precisely segment nodules from CT scan images. Bian et al. (4) proposed a dynamic multiscale weight optimisation network for getting spatial features inside CT scan images. Zhang et al.(19) introduced a multitask learning where a U-Net-based segmentation and a ResNet for classification. To improve the segmentation, feedback is performed on the classification result. Santone et al. (14) proposed a real time YOLO based model for segmentation. This model could capture small nodules properly. Hajim et al.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 5th Muslims in ML Workshop at NeurIPS'25.

(3) developed SEDARU-Net, a Squeeze-Excitation Dilated Attention-based Residual U-Net with intensity normalisation and YOLOv3 preprocessing, to improve lung nodule segmentation across diverse types and sizes, achieving high accuracy on the LUNA16 dataset.Agnes et al. (1) proposed a modified swinUnet where a Canny edge detector is used to highlight edges. However, the edge detector is used in the input image and aggregated with the final feature maps from swinUnet. So, the impact of the module is not very significant. Hou et al. (8) also used an attention mechanism to give importance to nodule edges, but the model is computationally expensive.

Although advancements have been made, several limitations remain: significant computational overhead, insufficient attention to refining nodule edges, and a lack of focus on segmentation model explainability.

To address these issues, a modified U-Net–based, edge-aware, Sobel-integrated model is proposed. This model demonstrated competitive performance compared to three baseline models and other state-of-the-art architectures. Furthermore, four XAI techniques were applied to enhance the trustworthiness of the model.

2 Method

This research begins with data acquisition, followed by preprocessing and data splitting, then proceeds with model training and testing using various evaluation metrics, and finally examines model transparency through several XAI techniques.

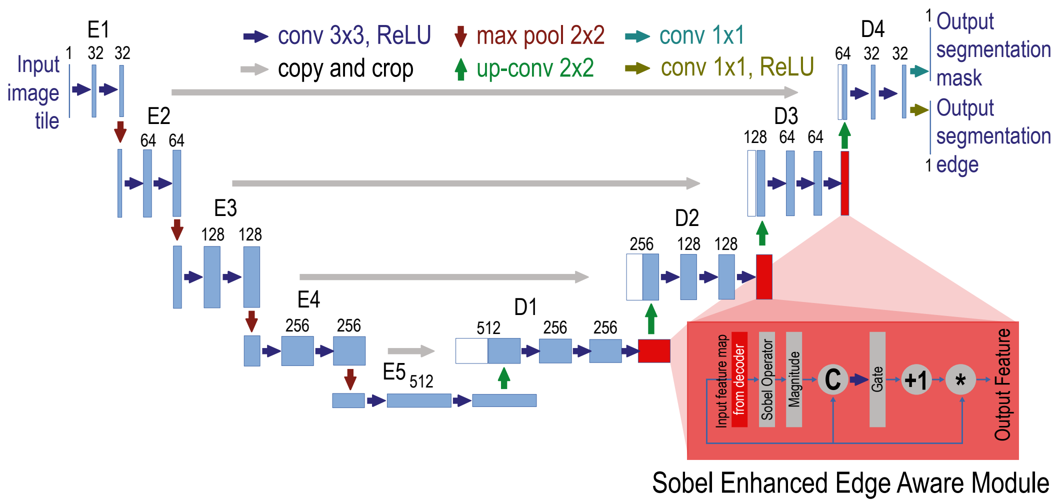

Figure 1: Modified U-Net with encoder (E1–E5), bottleneck (E5), and decoder (D1–D5). Each decoder output passes through an edge-aware block using Sobel (S), magnitude, concatenation (C), and gating, producing upsampled features for the final mask and edge prediction.

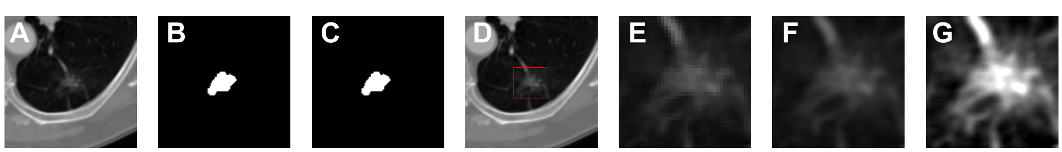

Figure 2: The Preprocessing pipeline: (a) Original grayscale, (b) Mask, (c) Binarised, (d) Nodule-centred square crop, (e) Cropped, (f) Resized 256×256, (g) Percentile clipping 2–98%.

The LIDC-IDRI dataset was used. The 2D slices with masks annotated by four radiologists were provided, and the dataset was collected from Kaggle, which contains some empty masks. After excluding empty masks, one valid image–mask pair was randomly selected per case, yielding 15,548 nodules from 875 patients. Each slice was converted to grayscale, masks were binarised, and lesion-centred crops with margins were resized to 256×256 (bilinear for images, nearest-neighbour for masks). Intensities were percentile-clipped to [0,1] for normalization. The dataset was split into five folds for cross-validation, with each fold containing about 175 patients, ensuring fair and unbiased evaluation. Figure 2 gives each preprocessing output of the dataset.

2.1 Proposed Model Architecture

Figure 1 shows the proposed model architecture and edge aware module . The proposed model is an **edge-aware U-Net** that consists of three main parts: an encoder, a bottleneck, and a decoder augmented with Sobel Edge Attention (SEA) modules.
The encoder extracts hierarchical features from the input image while progressively reducing the spatial resolution. At each stage, the feature maps are downsampled using max-pooling and refined by a DoubleConv block (two successive Conv–BN–ReLU operations), i.e., $E_{\ell+1} = \phi(\mathrm{BN}(W_{\ell,2} * \phi(\mathrm{BN}(W_{\ell,1} * \mathrm{Pool}(E_\ell)))))$, where E_ℓ is the feature map at stage ℓ, Pool denotes 2×2 max pooling, BN is batch normalization, $*$ is convolution, and ϕ is the ReLU activation.
At the lowest resolution, the bottleneck captures global semantic context as $B = \phi(\mathrm{BN}(W_{b,2} * \phi(\mathrm{BN}(W_{b,1} * E_L))))$, where E_L is the last encoder feature.
The decoder restores spatial detail by progressively upsampling features, concatenating them with corresponding encoder skip connections, and applying DoubleConv blocks, i.e., $\hat{D}_k = \mathrm{Up}(D_{k+1}),\ U_k = [\hat{D}_k, E_k],\ F_k = \phi(\mathrm{BN}(W_{k,2} * \phi(\mathrm{BN}(W_{k,1} * U_k)))))$.
Each decoder stage is followed by an SEA module, which emphasizes boundary information. For a feature map F, Sobel filters S_x, S_y (7) are applied channel-wise to compute gradient magnitude $M^c = \sqrt{(F^c * S_x)^2 + (F^c * S_y)^2}$, and a learned gate $G = \sigma(W_g * [F, M])$ is predicted to boost edge regions, producing $D = F \odot (1 + G)$.
Finally, the last decoder output D_4 is fed into two heads: a 1×1 convolution for segmentation logits $\hat{Y}$, and an auxiliary edge head for edge logits $\hat{E}$.

2.1.1 Loss Function

The architecture is trained with a multi-task loss that supervises both the segmentation and edge predictions. The segmentation branch is optimized with a combination of weighted binary cross-entropy and Dice loss (6), $\mathcal{L}_{\text{seg}} = \mathrm{BCE}(\sigma(\hat{Y}), y) + 0.5 \times (1 - \frac{2\sum(p \cdot y)}{\sum p + \sum y + \epsilon}))$, where $p = \sigma(\hat{Y})$ are predicted probabilities and y is the ground-truth mask. The auxiliary edge branch is supervised with binary cross-entropy against boundary labels y_{edge}, i.e., $\mathcal{L}_{\text{edge}} = \mathrm{BCE}(\sigma(\hat{E}), y_{\text{edge}})$.

The total objective combines both segmentation and edge terms as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{edge}} \mathcal{L}_{\text{edge}}, \tag{1}$$

where λ_{edge} controls the contribution of the edge loss. The effect of varying λ_{edge} is analyzed through an ablation study, as shown in Table 3.

so that the decoder is guided to predict accurate nodule regions while the edge branch and SEA modules are explicitly supervised to sharpen boundary localization.

2.2 Experimental Settings, Evaluation Metrics and Explainability

Training used a batch size of 8, for 60 epochs with patience 10, optimized with AdamW (lr 1e-3, weight decay 1e-5, betas 0.9/0.999, eps 1e-8). The model had 32 base channels and ran on an NVIDIA RTX 3090 GPU. The models' results are evaluated by six metrics: Dice score (DSC), Intersection over union (IoU), Accuracy, Sensitivity, Specificity and Harmonic Mean (12). Four XAI techniques: Grad-CAM (15), Score-CAM (17), Layer-CAM and Counterfactual-XAI (11) are used to highlight regions that were used for edge prediction. The source code is available at: `https://github.com/brai-acslab/edge-aware-unet`.

3 Results and Discussion

Table 1 compares our model with three baselines. It outperforms others in all metrics except specificity and IoU. While accuracy matches UNet++ (0.90), our model has a much lower standard deviation (0.005 vs. 0.24), indicating stronger generalizability. Table 2 compares the computational efficiency of different models. The proposed model achieves a balanced trade-off between speed and

Table 1: Comparison of segmentation performance across models (Mean ± Std).

Model	DSC	IoU	Accuracy	Sensitivity	Specificity	Harmonic Mean
U-Net	0.81 ± 0.33	0.78 ± 0.31	0.83 ± 0.39	0.79 ± 0.35	0.92 ± 0.22	0.85 ± 0.26
ResUNet++	0.86 ± 0.41	0.75 ± 0.51	0.83 ± 0.31	0.72 ± 0.37	**0.95 ± 0.29**	0.82 ± 0.33
U-Net++	0.88 ± 0.27	**0.79 ± 0.31**	0.90 ± 0.24	0.87 ± 0.27	0.94 ± 0.21	0.90 ± 0.24
Proposed	**0.90 ± 0.005**	0.74 ± 0.006	**0.90 ± 0.005**	**0.92 ± 0.017**	0.89 ± 0.014	**0.90 ± 0.015**

Table 2: Comparison of model complexity (batch size = 1) on an RTX 3090 for image size of 256×256 .

Model	Params (M)	GFLOPs	Latency(ms / image)
UNet	3.35	15.59	2.18
UNet++	6.21	66.82	5.02
ResUNet++	4.17	27.13	3.87
Proposed	**3.40**	**17.29**	**3.41**

complexity—requiring only 3.4 M parameters and 17.3 GFLOPs, with an inference time of 3.41 ms per image. This makes it nearly as lightweight as U-Net but substantially faster and more efficient than U-Net++ and ResUNet++.

Table 3: Ablation study of the loss function for a single fold by changing the coefficient λ_{edge} of the edge loss in Eq. 1.

λ	Accuracy	Dice	IoU	Sensitivity	Specificity	Harmonic Mean
0.25	0.8894	0.8384	0.7218	0.9507	0.8629	0.9045
0.50	**0.9034**	**0.8508**	**0.7403**	0.9123	**0.8995**	0.9059
0.75	0.8859	0.8361	0.7184	**0.9639**	0.8522	0.9058
1.00	0.8989	0.8485	0.7368	0.9375	0.8822	**0.9094**

Table 3 presents the ablation study on the edge-loss weighting coefficient λ_{edge}. The results show that setting $\lambda = 0.5$ yields the best overall balance across all metrics, achieving the highest Accuracy, Dice, IoU, and Specificity. Lower or higher values of λ either overemphasize edge information or reduce its influence, slightly degrading segmentation quality. Thus, $\lambda = 0.5$ was selected as the optimal configuration for the proposed model.

Table 4: Comparison of segmentation performance across different SOTA.

Method	DSC	IoU	Accuracy	Sensitivity	Specificity
Agnes et al (2)	–	–	–	0.726	–
Kido et al. (9)	–	–	–	0.738	–
Usman et al. (16)	–	–	–	–	–
SAR-UNet (18)	0.8846	0.8971	0.8760	0.8745	0.8810
Attention-UNet (13)	0.9190	0.9179	0.9071	0.9063	0.9190
Swin-UNet (5)	0.8674	0.8697	0.8518	0.8501	0.8670
Proposed	**0.90**	**0.74**	**0.90**	**0.92**	**0.89**

Table 4 presents the performance comparison with other state-of-the-art architectures. Our model outperforms all others across all metrics, except for Attention U-Net (13), which is only about 1% higher. However, Attention U-Net requires substantially more computational resources, whereas our model is a simple modified U-Net, demonstrating superior computational efficiency.

3.1 Intrepretability

Figure 3 illustrates the predicted edges, segmentation masks, and heatmaps generated by four XAI techniques. Grad-CAM clearly highlights the nodule boundaries, with Score-CAM and Layer-CAM showing similar emphasis on edge regions. Counterfactual-XAI, despite slight perturbations in the input, produced stable edge predictions, underscoring the model's transparency and robustness.

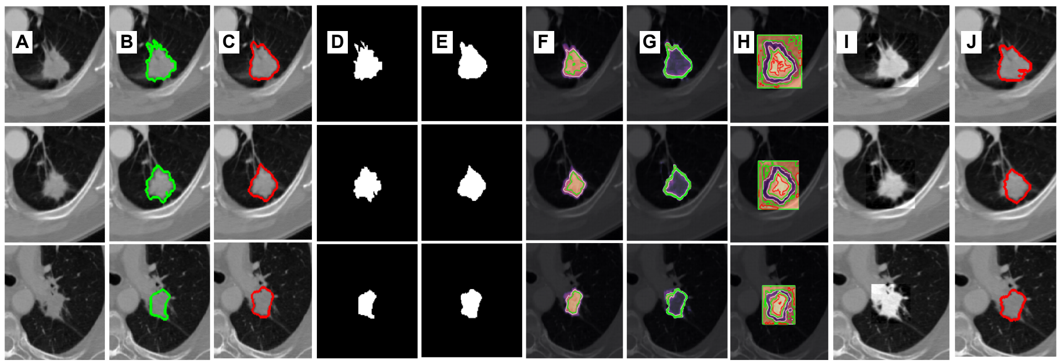

Figure 3: Heatmap generation of all XAI techniques: (a) Raw image, (b) GT edges, (c) Predicted edges, (d) GT mask, (e) Predicted mask, (f) Grad-CAM, (g) Layer-CAM, (h) Score-CAM, (i) Counterfactual image, (j) Counterfactual predicted edges.

4 Limitations and Future Work

Although this work is effective and explainable, there are several limitations: the dataset was already 2D-sliced. The actual 3D images can give more insight. Moreover, there are other metrics which were not used for the evaluation. Based on the explanations from XAI techniques, the model is not retrained to improve performance. Traditional Unet is simple; adding an edge-aware module to other powerful segmentation models, like transformers, can be more effective. Moreover, future work will focus on incorporating zero-shot and fine-tuned predictions using large foundation models to enhance generalization and diagnostic robustness .

5 Conclusion

The presented research demonstrates that the proposed model can accurately predict both segmentation masks and nodule edges from CT scans. Despite its simplicity, the model outperforms several state-of-the-art architectures. With only 3.4 M parameters, significantly fewer than other U-Net variants, it is computationally efficient and suitable for clinical deployment. Moreover, the XAI techniques showed that the model is robust and can provide clinicians with greater confidence in its predictions.

References

[1] S Akila Agnes, A Arun Solomon, K Karthick, Mejdl S Safran, and Sultan Alfarhood. Edge-enhanced feature pyramid swinunet: Advanced segmentation of lung nodules in ct images. *IET Image Process.*, 19(1), 2025.

[2] Sundaresan A Agnes and Jeevanayagam Anitha. Efficient multiscale fully convolutional unet model for segmentation of 3d lung nodule from ct image. *Journal of Medical Imaging*, 9(5):052402–052402, 2022.

[3] Dhafer Alhajim, Karim Ansari-Asl, Gholamreza Akbarizadeh, and Mehdi Naderi Soorki. Improved lung nodule segmentation with a squeeze excitation dilated attention based residual unet. *Scientific Reports*, 15(1):3770, 2025.

[4] Xinjun Bian, Huan Lin, Yumeng Wang, Lingqiao Li, Zhenbing Liu, Huadeng Wang, Zhenwei Shi, Yi Qian, Rushi Lan, Xipeng Pan, et al. Federated cross-source learning for lung nodule

segmentation with data characteristic-aware weight optimization. *Pattern Recognition*, page 112396, 2025.

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[6] Adrian Galdran, Gustavo Carneiro, and Miguel Ángel González Ballester. On the optimal combination of cross-entropy and soft dice losses for lesion segmentation with out-of-distribution robustness, 2022. URL: `https://arxiv.org/abs/2209.06078`, `arXiv:2209.06078`.

[7] Wenshuo Gao, Xiaoguang Zhang, Lei Yang, and Huizhong Liu. An improved sobel edge detection. In *2010 3rd International Conference on Computer Science and Information Technology*, volume 5, pages 67–71, 2010. `doi:10.1109/ICCSIT.2010.5563693`.

[8] Jiachen Hou, Chuan Yan, Ru Li, Qingnan Huang, Xiangsuo Fan, and Fangyu Lin. Lung nodule segmentation algorithm with smr-unet. *IEEE Access*, 11:34319–34331, 2023. `doi:10.1109/ACCESS.2023.3264789`.

[9] Shoji Kido, Shunske Kidera, Yasushi Hirano, Shingo Mabu, Tohru Kamiya, Nobuyuki Tanaka, Yuki Suzuki, Masahiro Yanagawa, and Noriyuki Tomiyama. Segmentation of lung nodules on ct images using a nested three-dimensional fully connected convolutional network. *Frontiers in artificial intelligence*, 5:782225, 2022.

[10] Van-Linh Le and Olivier Saut. Rrc-unet 3d for lung tumor segmentation from ct scans of non-small cell lung cancer patients. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2316–2325, 2023.

[11] Ioannis E. Livieris, Emmanuel Pintelas, Niki Kiriakidou, and Panagiotis Pintelas. Explainable image similarity: Integrating siamese networks and grad-cam. *Journal of Imaging*, 9(10), 2023. URL: `https://www.mdpi.com/2313-433X/9/10/224`, `doi:10.3390/jimaging9100224`.

[12] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation, 2022. URL: `https://arxiv.org/abs/2202.05273`, `arXiv:2202.05273`.

[13] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[14] Antonella Santone, Francesco Mercaldo, and Luca Brunese. A method for real-time lung nodule instance segmentation using deep learning. *Life*, 14(9):1192, 2024.

[15] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL: `http://arxiv.org/abs/1610.02391`, `arXiv:1610.02391`.

[16] Muhammad Usman and Yeong-Gil Shin. Deha-net: A dual-encoder-based hard attention network with an adaptive roi mechanism for lung nodule segmentation. *Sensors*, 23(4):1989, 2023.

[17] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks, 2020. URL: `https://arxiv.org/abs/1910.01279`, `arXiv:1910.01279`.

[18] Jinke Wang, Peiqing Lv, Haiying Wang, and Changfa Shi. Sar-u-net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual u-net for automatic liver segmentation in computed tomography. *Computer Methods and Programs in Biomedicine*, 208:106268, 2021.

[19] Jiasen Zhang, Mingrui Yang, Weihong Guo, Brian A Xavier, Michael Bolen, and Xiaojuan Li. Detection-guided deep learning-based model with spatial regularization for lung nodule segmentation. *Quantitative Imaging in Medicine and Surgery*, 15(5):4204, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in abstract is reflected in Introduction and Experiments section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: There is a separate Limitations section provided in the paper

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our approach focuses on enhancing the practical results and no theoretical aspects are presented. Hence, full set of assumptions and complete proof are not applicable

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 (Experiments) provides necessary information to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides a reference to the data used in the experiment. Additionally, a GitHub link is provided to access the code base for the reproduction of the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies training, testing and validation splits and provides exact files used in training and validation as CSV files in the released code base.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Section 3 presents experiments that support the main claims of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.1 details information on the compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The dataset used in the papaer was already anonymized prior to acquiring it. Hence, there are no personally identifiable information in the dataset.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper does not have direct societal impacts as it is designed for sign language recognition tasks.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The release code has no risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The data used in the study is credited with appropriate citation ((**?**))

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code used in the study has been released through GitHub repository and is mentioned in Section 3.1

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The study involves no crowdsourcing experiments and research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM was used for editing the initial content written by authors.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.