RASALoRE: Region-Aware Spatial Attention with Location-based Random Embeddings for Weakly Supervised Anomaly Detection in Brain MRI Scans

Bheeshm Sharma[1], *Karthikeyan Jaganathan*[2], *Balamurugan Palaniappan*[3]

[1,3]*Department of Industrial Engineering & Operations Research (IEOR), IIT Bombay, India*
[2]*Department of Energy Science and Engineering (DESE), IIT Bombay, India*

{[1]bheeshmsharma, [2]karthikeyanj, [3]balamurugan.palaniappan}@iitb.ac.in

BMVC 2025

Introduction

Anomaly detection in brain MRI scans is a widely recognized task, helpful in timely identification and treatment of related illnesses, but becomes challenging due to limited availability of labeled data with accurate pixel-wise annotations.

Key Challenges:

- Intricate complexity of brain anatomy
- Limited labeled data with accurate pixel-wise annotations

Our Contributions:

- A novel two-stage weakly supervised anomaly detection (WSAD) framework that operates using only slice-level labels.
- A prompt-tuning strategy to generate high-quality pseudo-masks from slice level supervision.
- A **Region-Aware Spatial Attention** (RASA) mechanism guided by **Location-based Random Embeddings** (LoRE) to effectively capture local contextual dependencies.
- Our method achieves superior performance while significantly reducing the number of model parameters.

Stage-1: Discriminative Dual Prompt Tuning (DDPT)

In Stage-1, we propose Discriminative Dual Prompt Tuning (DDPT), a classification-driven framework that generates coarse anomaly segmentation maps under weak slice-level supervision. The model jointly optimizes learnable vision and text prompts to classify MRI slices and helps derive attention maps for pseudo-mask generation.

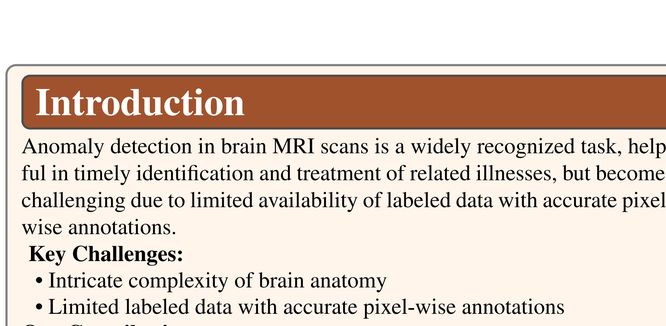

Figure 1: Overview of Discriminative Dual Prompt Tuning (DDPT).

Key Components

1. Learnable Text Prompts:

- Structured as: $t = [V]_1 \dots [V]_{\mathrm{M}/2}\,[\mathrm{CLASS}]\,[V]_{\mathrm{M}/2+1} \dots [V]_{\mathrm{M}}$
- Here, $[V]_i$ are learnable tokens and [CLASS] represents either `healthy` or `unhealthy`.

2. Vision Transformer (ViT) with Visual Prompts:

- ViT backbone weights are frozen; only prompt parameters are trainable.
- Class-Aware Visual Prompt Tuning (CAVPT) refines the visual tokens.

3. Overall Loss Function:

$$\mathcal{L}_{\text{total}} = \eta \mathcal{L}_{\text{ce}}^{\text{ca}} + \mathcal{L}_{\text{ce}}$$

where $\mathcal{L}_{\text{ce}}$ is the standard cross-entropy between predicted and true labels in ViT, $\mathcal{L}_{\text{ce}}^{\text{ca}}$ is an auxiliary cross-entropy loss applied to the output of the class-aware visual prompt generator in CAVPT, and η controls the relative weighing of both terms.

Output: DDPT provides pseudo weak masks M_{DDPT} for anomalous regions.

Stage 2: RASALoRE Architecture

In Stage-2, we propose RASALoRE, a segmentation network that leverages fixed location-based random embeddings (LoRE) within a Region-Aware Spatial Attention (RASA) module. Guided by the pseudo-masks generated in Stage-1, it enables precise anomaly localization under weak supervision.

Location-based Random Embeddings (LoRE)

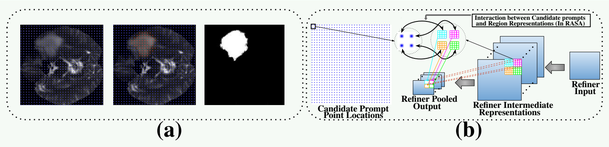

Figure 2: (a) Left: Candidate prompt point locations (in blue) overlaid as grid on input image, center: point activation mask (red denoting active and blue denoting inactive points) overlaid on input image, right: weak anomaly mask corresponding to input image. (b) Refiner Module.

- We use $\sqrt{k} \times \sqrt{k}$ grid of k evenly spaced **Candidate Prompt Points (CPPs)**.
- We generates fixed, non-learnable random embeddings based on sinusoidal transformations.
- Designed to be independent of dataset-specific biases.
- These random embeddings $E_{\text{cpp}} \in \mathbb{R}^{k \times d}$ remain fixed during both training and inference.

Refiner Module: Processes input X to output $R_\rho(X) \in \mathbb{R}^{\sqrt{k} \times \sqrt{k} \times d}$, where each output pixel corresponds to a CPP location.

Region-Aware Spatial Attention (RASA):

Region-Aware Spatial Attention (RASA):

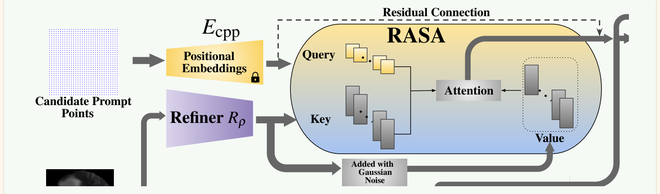

Figure 3: Overview of RASA

RASA enables interaction between location embeddings and spatial information:

- Multi-head attention: $\text{MHA}_{\text{RASA}}(Q, K, V)$
- **Query:** $Q = E_{\text{cpp}}$, **Key:** $K = R_\rho(X)$ (refiner output)
- **Value:** $V = R_\rho(X) + \epsilon$ (with Gaussian noise for robustness)
- **Output:** Enriched Spatial Point Embeddings $\xi_{\text{ESPE}} \in \mathbb{R}^{k \times d}$

Mask Decoder

The decoder performs attention between enriched embeddings ξ_{ESPE} and image features:

- Image encoder U_ϱ provides intermediate feature representations
- $\text{MHA}_{\text{Dec}}(Q = \xi_{\text{ESPE}}, K = U_\varrho(X), V = U_\varrho(X) + \epsilon)$
- Upsampling produces final anomaly mask $M_{\text{ANO}} \in \mathbb{R}^{h \times w}$

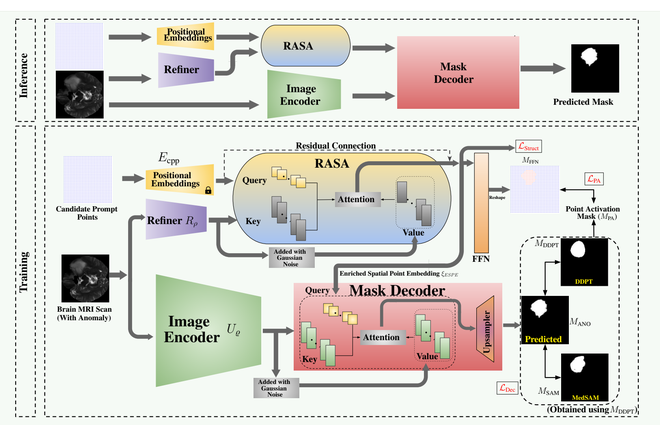

Figure 4: Overview of RASALoRE Architecture.

Training Loss Functions

1. Decoder Loss ($\mathcal{L}_{\text{Dec}}$)

$$\begin{aligned}\mathcal{L}_{\text{Dec}} = \text{ELDice}\left(M_{ANO}, G_\sigma(M_{DDPT})\right) + \gamma \cdot \text{ELDice}\left(M_{ANO}, G_\sigma^{-1}(M_{SAM})\right) \\ + \frac{\alpha}{p} \cdot (M_{ANO} \odot (1 - M_{DDPT})) \\ + \beta \cdot \text{ELDice}\left((1 - M_{ANO}) \odot (1 - M_{DDPT}), G_\sigma(1 - M_{DDPT})\right)\end{aligned}$$

Key Features:

- Uses DDPT masks with Gaussian filter G_σ (center-weighted)
- Uses MedSAM masks with inverse Gaussian G_σ^{-1} (boundary-weighted)
- Weight $\gamma = \text{Dice}(M_{SAM}, M_{DDPT})$ ensures quality
- Controls false positives through third and fourth terms

2. Point Activation Loss ($\mathcal{L}_{\text{PA}}$)

$$\mathcal{L}_{\text{PA}} = \text{ELDice}\left(M_{\text{FFN}}, M_{PA}\right)$$

Compares FFN output with point activation mask from DDPT.

3. Structural Loss ($\mathcal{L}_{\text{Struct}}$)

$$\mathcal{L}_{\text{Struct}} = \text{MSE}\left(\xi_{\text{ESPE}}^{A}, \mathbf{1}\right) + \text{MSE}\left(\xi_{\text{ESPE}}^{IA}, -\mathbf{1}\right)$$

Enforces similarity among embeddings for active/inactive points.

Total Loss

$$\mathcal{L} = \mathcal{L}_{\text{Dec}} + \mathcal{L}_{\text{PA}} + \mathcal{L}_{\text{Struct}}$$

Quantitative Results

Method	BraTS20		BraTS21		BraTS23		MSD	
	Dice	AUPRC	Dice	AUPRC	Dice	AUPRC	Dice	AUPRC
AE	14.26	10.23	11.83	8.01	17.09	7.41	14.96	7.07
DAE	21.33	18.89	14.38	14.59	34.16	21.18	32.13	20.77
VQVAE	17.27	12.04	25.69	17.67	20.67	38.48	19.44	33.78
TS-CAM	6.13	7.92	6.74	8.35	9.13	9.36	7.81	8.47
CAE	26.36	17.48	23.82	14.20	46.98	60.11	27.96	18.64
LA-GAN	34.14	28.48	42.75	38.82	40.57	43.65	33.63	27.70
AME-CAM	52.22	37.39	50.43	37.85	39.19	26.47	51.91	40.34
AnoFPDM	37.18	38.78	41.83	47.89	49.28	57.04	43.42	50.08
Yoo et al. (T2)	22.76	13.12	11.94	8.64	12.20	8.90	12.09	8.79
Yoo et al. (All)	49.91	38.64	63.33	50.28	23.41	12.99	47.81	35.93
DDPT	61.53	46.89	51.72	35.79	48.59	31.66	48.71	33.87
M2+DDPT(p)	32.57	24.10	34.73	25.55	48.52	39.93	43.43	34.75
M2+DDPT(b)	35.58	25.44	37.24	26.54	53.54	43.15	45.90	35.78
M+DDPT(p)	37.66	26.29	43.44	29.49	39.22	25.39	38.08	25.46
M+DDPT(b)	43.44	33.36	51.19	36.54	50.40	33.66	46.46	33.02
R. w/o MedSAM	69.80	73.06	68.87	74.26	74.22	80.70	61.34	67.08
RASALoRE	70.57	74.74	70.85	75.05	70.79	71.18	61.37	63.71

Key Observations:

- Classical reconstruction-based models (AE, DAE, VQVAE) achieve relatively low scores and CAM-based methods (CAE, LA-GAN, AME-CAM) improve performance but exhibit considerable fluctuations between datasets.
- RASALoRE demonstrates consistently strong and stable performance across all datasets (BraTS20, BraTS21, BraTS23, MSD), achieving significant improvements in key metrics such as Dice and AUPRC.
- R. Without MedSAM shows that DDPT-generated weak masks alone suffice for strong segmentation, nearly matching full RASALoRE performance.

Qualitative Results

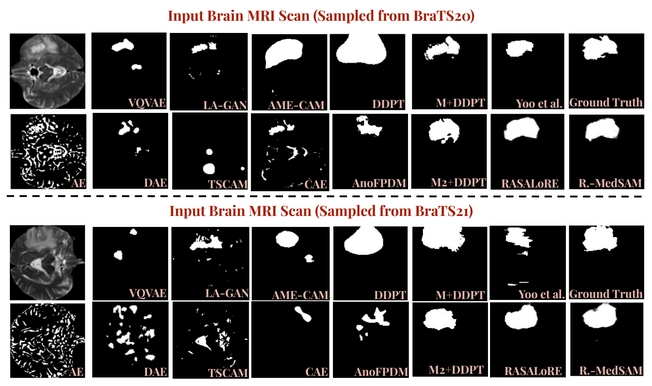

Figure 5: Qualitative comparison of predicted anomaly masks.

Visual Analysis via Qualitative Results:

- **Reconstruction methods:** Blurred boundaries, incomplete segmentations
- **CAM methods:** Partial improvements but miss fine details
- **RASALoRE:** Comparatively sharp boundaries, accurate localization

Multimodality RASALoRE

Extended to multiple MRI modalities using T2 as bridge.

Dataset	Dice Score (%)				AUPRC (%)			
	T1	T2	T1ce	Flair	T1	T2	T1ce	Flair
BraTS20	65.13	71.82	66.77	72.42	66.90	77.08	68.86	75.62
BraTS21	63.24	68.57	67.65	69.53	62.87	73.55	67.67	74.23
BraTS23	54.24	61.17	57.60	63.18	53.46	61.76	54.34	63.02
MSD	56.19	67.31	61.18	69.20	60.66	73.63	64.04	74.59

Key Features:

- Modality-specific RASA modules with shared encoder/decoder
- Bridge alignment loss: $\mathcal{L}_{\text{align}} = \sum_j \|\xi_{\text{ESPE}}^{(j)} - \xi_{\text{ESPE}}^{\text{bridge}}\|_2^2$
- T1 and T1ce modalities achieve competitive performance
- Parameter-efficient: same inference cost across modalities

Conclusions

- **RASALoRE**, a weakly supervised anomaly detection technique useful for anomaly segmentation in brain MRI scans, when ground-truth pixel-level annotations are unavailable.
- RASALoRE uses fixed candidate prompt point locations whose location-based random embeddings interact with suitable image-level intermediate feature representations, to provide sufficiently rich region-aware embeddings that elicit localized anomaly information from MRI scan images.
- We introduce **Discriminative Dual Prompt Tuning (DDPT)**, a novel vision-language prompt-tuning strategy that generates pseudo-masks to supervise RASALoRE training effectively.
- Our results demonstrate the strong and efficient detection capabilities of RASALoRE across diverse BraTS-type datasets, using fewer than approximately 7.8 million parameters.
- We successfully extend RASALoRE to **multimodal MRI** (T1, T1ce, Flair) using T2 as a bridge modality.

Acknowledgments

We gratefully acknowledge Technocraft Centre of Applied Artificial Intelligence (TCAAI), IIT Bombay for support through generous funding.

References

1. Zhou et al., *Learning deep features for discriminative localization.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
2. Chen et al., *Ame-CAM: Attentive multiple-exit CAM for weakly supervised segmentation on MRI brain tumor.* In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 173–182. Springer, 2023.
3. Xie et al., *A weakly supervised and globally explainable learning framework for brain tumor segmentation.* In 2024 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE, 2024.
4. Che et al., *AnoFPDM: Anomaly detection with forward process of diffusion models for brain MRI.* In Proceedings of the Winter Conference on Applications of Computer Vision, pp. 1113–1122, 2025.
5. Ma et al., *Segment anything in medical images.* *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z.
6. Zhou et al., *Learning to prompt for vision-language models.* *International Journal of Computer Vision*, 130(9):2337–2348, 2022.
7. Xing et al., *Dual modality prompt tuning for vision-language pre-trained model.* *IEEE Transactions on Multimedia*, 2023.

Full Paper and Code

Paper

Code