DEFUSE-MS: Deformation Field-Guided Spatiotemporal Graph-Based Framework for Multiple Sclerosis New Lesion Detection

Mostafa Salem[1,2] ✉[0000−0001−9915−8390], Salma Hassan[1][0009−0006−5498−8319], Vijay Papineni[3][0000−0002−6162−3290], Ayman Elsayed[3][0000−0003−3381−066X], and Mohammad Yaqub[1][0000−0001−6896−1105]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{Mostafa.Salem, Salma.Hassan, Mohammad.Yaqub}@mbzuai.ac.ae
[2] Department of Computer Science, Faculty of Computers and Information, Assiut University, Assiut, Egypt
[3] Sheikh Shakhbout Medical City, Abu Dhabi, UAE
rkpvijay@gmail.com, ayaahmed@ssmc.ae

Abstract. Longitudinal magnetic resonance imaging (MRI) is essential for diagnosing and monitoring multiple sclerosis (MS), a chronic central nervous system disorder. Tracking brain lesion evolution over time is essential for predicting MS progression, yet this process is time-consuming and subject to intra- and interobserver variability. While deep learning models such as convolutional neural networks (CNNs) and vision transformers (ViTs) have been applied to lesion detection, they often struggle to fully capture spatial, structural and temporal relationships. Vision graph neural networks (ViGs) present a novel approach with the potential to improve performance in these tasks by effectively capturing relational and structural information. We introduce DEFUSE-MS, a Deformation Field-Guided Spatiotemporal ViG-Based Framework for detecting MS new T2-weighted lesions. The framework features a Heterogeneous Spatiotemporal Graph Module (HSTGM), which functions as both an encoder and decoder. Evaluated on the MSSEG-II dataset, DEFUSE-MS achieves state-of-the-art performance with a lesion detection F1 score of 0.65, sensitivity (SensL) of 0.74, positive predictive value (PPVL) of 0.65, and a mean segmentation Dice score of 0.55, outperforming the state-of-the-art methods. These results highlight DEFUSE-MS's efficacy in MS new lesion detection. The code is available at https://github.com/BioMedIA-MBZUAI/DEFUSE-MS.

Keywords: Brain · MRI · Multiple Sclerosis · Deep Learning · Spatiotemporal Learning · Vision Graph Neural Network

1 Introduction

Multiple Sclerosis (MS) ranks among the leading causes of neurological impairment in adults in the present day, affecting nearly 3 million people worldwide

[9]. Early diagnosis is crucial for timely intervention, which can help slow disease progression. Magnetic resonance imaging (MRI) plays a key role in both diagnosing and monitoring MS, with the appearance of new T2-weighted (T2-w) lesions serving as a critical predictor of disease activity [26]. Manually tracking these changes over time is labor-intensive, error-prone, and subject to intra- and interobserver variability [1]. As a result, there is a growing demand for automated methods that offer fast, reliable, and consistent assessments, particularly for measuring lesion volumetric changes over time.

MS lesion analysis encompasses both lesion detection in single MRI scans and change detection across longitudinal scans. These lesions manifest through tissue transformation (intensity shifts) and deformation (changes in adjacent tissue) [14]. Change detection methodologies typically follow two approaches: intensity-based comparisons [23] and deformation-based analyses, which use non-rigid registration deformation fields (DFs) for new T2-w lesion identification [2]. Hybrid approaches have emerged that combine both methodologies to improve accuracy [20,22]. In 2021, the MSSEG-II challenge, organized by MICCAI, was launched to compare automated solutions for the detection of MS new lesions appearing at the second time point of two FLAIR images of the patient [5]. Recent studies using this dataset have reported their performance exclusively through cross-validation on the training set, without evaluating on the test cases [29,24]. While convolutional neural networks (CNNs), particularly U-Net [19], have revolutionized medical image segmentation, their local nature limits global context modeling [28]. Transformer-based architectures such as TransUNet [4] and UNETER [11] address this limitation but require extensive training data [8]. However, vision graph neural networks (ViGs) offer a promising alternative by representing images as graphs, effectively combining CNN's local feature extraction with Transformer-like global context modeling through graph neural networks (GNNs) [10].

We introduce DEFUSE-MS, Deformation Field-Guided Spatio-temporal ViG-based framework for MS new T2-w Lesion Detection. We frame the problem of detecting MS new lesions as a heterogeneous spatiotemporal GNN employing an encoder-decoder architecture. The proposed model utilizes baseline and follow-up MRI scans to construct a heterogeneous spatiotemporal graph, where graph nodes are connected by two types of spatial edge and one type of temporal edge. Temporal edges are further augmented with learned DF embeddings as temporal edge attributes, capturing lesion evolution patterns that are clinically relevant for assessing disease progression. This approach effectively captures spatial relationships within individual scans and temporal dynamics between the two-time points, allowing a comprehensive analysis of lesion evolution. To the best of our knowledge, this is the first study to leverage ViGs for 3D medical image segmentation, marking a significant advancement in the field. Our key contributions include: (1) introducing a novel formulation for MS new lesion detection in 3D MRI as a heterogeneous **spatiotemporal GNN** within an encoder-decoder framework; (2) integrating **learned DF embeddings as edge attributes** in the heterogeneous spatiotemporal graph, introducing a robust representation

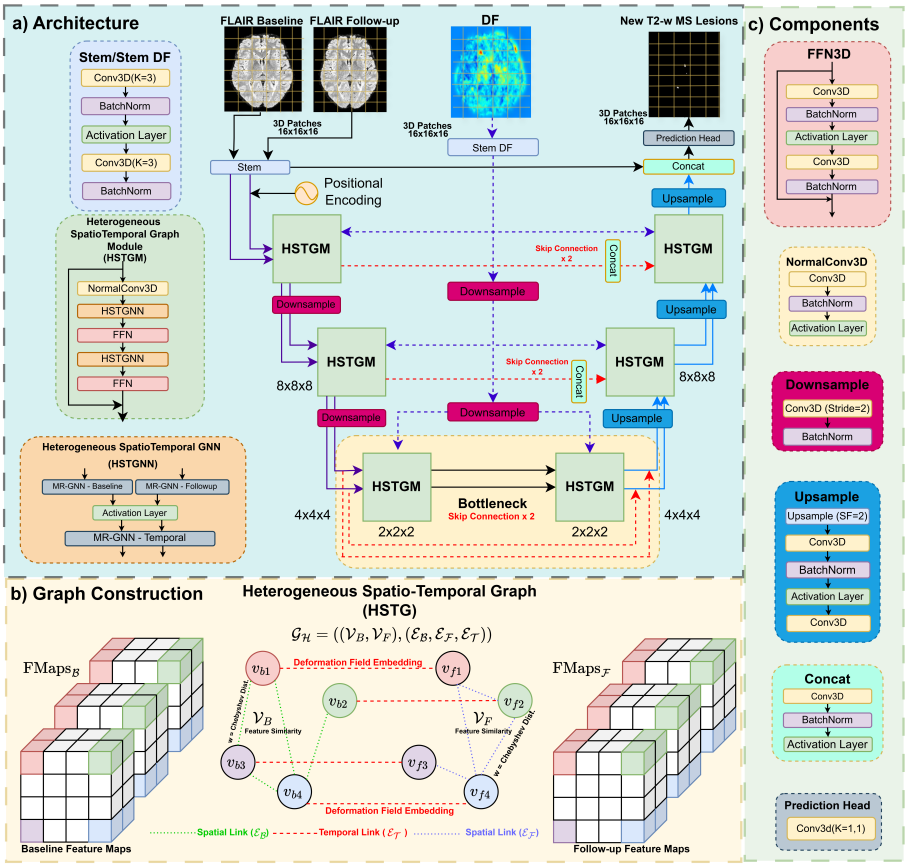

Fig. 1: Overview of the DEFUSE-MS framework for MS new T2-w lesion detection. **(a) DEFUSE-MS Architecture:** The proposed network consists of a U-shaped encoder-decoder model. The inputs include baseline and follow-up images along with the DF, which nonlinearly registers the baseline image to the follow-up image. **(b) Graph Construction:** Illustrates how the heterogeneous spatio-temporal graph is reconstructed from the baseline and follow-up feature maps. **(c) Components:** Details of the DEFUSE-MS components.

that captures the clinical manifestation of disease progression over time, and (3) achieving state-of-the-art performance on the **MSSEG-II** challenge test set.

2 Methodology

DEFUSE-MS is a 3D patch-wise U-shaped encoder-decoder model, as depicted in Fig. 1 (a). This model processes FLAIR modality images from baseline and follow-up scans along with the DF to generate the new T2-w lesion segmentation

mask. Its structure comprises an encoder, a bottleneck, a decoder, and skip connections. Key components include stem blocks (convolutional modules that generate initial feature maps, serving as the foundation for constructing graph node features and temporal edge attributes), a Heterogeneous Spatio-Temporal Graph Module (HSTGM), feedforward networks (FFNs), and downsample and upsample modules. The learned DF embeddings are shared between the encoder and decoder, facilitating effective spatial information integration.

Graph Construction. Image stem and DF stem generate output feature maps ($\text{FMaps}_{\mathcal{B}}$, $\text{FMaps}_{\mathcal{F}}$), and $\text{FMaps}_{\mathcal{DF}}$, respectively, all in $\mathbb{R}^{C\times D\times H\times W}$, to form the Heterogeneous Spatio-Temporal Graph (HSTG). In Fig. 1 (b), the baseline feature maps ($\text{FMaps}_{\mathcal{B}}$) and follow-up feature maps ($\text{FMaps}_{\mathcal{F}}$) are treated as unordered node sets: $\mathcal{V}_B = \{v_{b1}, v_{b2}, \ldots, v_{bn}\}$ and $\mathcal{V}_F = \{v_{f1}, v_{f2}, \ldots, v_{fn}\}$, each in $\mathbb{R}^{C\times DHW}$, respectively. These nodes form the vertices of the heterogeneous graph. For any baseline node $v_{bi} \in \mathbb{R}^C$ and follow-up node $v_{fi} \in \mathbb{R}^C$, $\mathcal{K}$ nearest neighbors, $\mathcal{N}(v_{bi})$ and $\mathcal{N}(v_{fi})$, are selected via Euclidean distance between features. The HSTG includes three edge types: 1) baseline spatial edges ($\mathcal{E}_{\mathcal{B}}$) link nearest neighbors in the baseline, 2) follow-up spatial edges ($\mathcal{E}_{\mathcal{F}}$) link nearest neighbors in the follow-up and 3) temporal edges ($\mathcal{E}_{\mathcal{T}}$) link corresponding nodes between baseline and follow-up. Spatial edges e_{s-ji} from v_j to v_i have attributes derived from the Chebyshev distance of their spatial coordinates and the Euclidean distance of their feature vectors ($e_{s-ji} \in \mathbb{R}^2$). Temporal edges e_{t-ii}, directed from v_{bi} to v_{fi}, adopt DF-derived learned embeddings ($e_{t-ii} \in \mathbb{R}^C$). The HSTG is expressed as $\mathcal{G}_{\mathcal{H}} = ((\mathcal{V}_B, \mathcal{V}_F), (\mathcal{E}_{\mathcal{B}}, \mathcal{E}_{\mathcal{F}}, \mathcal{E}_{\mathcal{T}}))$.

Heterogeneous Spatiotemporal Graph Module (HSTGM). The HSTGM serves as a core component of the DEFUSE-MS network, functioning as both the encoder and decoder block. It includes a 3D convolutional layer followed by two Heterogeneous Spatio-Temporal Graph Neural Network (HSTGNN) layers. Each HSTGNN layer is followed by an FFN to enhance feature transformation capacity and mitigate the over-smoothing effect. Each HSTGNN layer comprises three Max-Relative Graph Neural Networks (MR-GNNs): Two spatial MR-GNNs ($\mathcal{F}_{\mathcal{B}}$, $\mathcal{F}_{\mathcal{F}}$) independently aggregate and update the baseline and follow-up graphs ($\mathcal{G}_{\mathcal{B}}, \mathcal{G}_{\mathcal{F}}$) via spatial edges, and a temporal MR-GNN ($\mathcal{F}_{\mathcal{T}}$) that aggregates and updates features between timepoints via temporal edges.

$$\mathcal{G}'_{\mathcal{B}} = \mathcal{F}_{\mathcal{B}}(\mathcal{G}_{\mathcal{B}}, \mathcal{W}_{\mathcal{B}}), \quad \mathcal{G}'_{\mathcal{F}} = \mathcal{F}_{\mathcal{F}}(\mathcal{G}_{\mathcal{F}}, \mathcal{W}_{\mathcal{F}}) \quad \mathcal{G}'_{\mathcal{H}} = \mathcal{F}_{\mathcal{T}}((\mathcal{G}'_{\mathcal{B}}, \mathcal{G}'_{\mathcal{F}}), \mathcal{W}_{\mathcal{T}}), \quad (1)$$

where $\mathcal{W}_{\mathcal{B}}$, $\mathcal{W}_{\mathcal{F}}$ and $\mathcal{W}_{\mathcal{T}}$ are the learnable weights of the MR-GNNs $\mathcal{F}_{\mathcal{B}}$, $\mathcal{F}_{\mathcal{F}}$, $\mathcal{F}_{\mathcal{T}}$, respectively.

For a graph $\mathcal{G} = \mathcal{G}(\mathcal{X}, \mathcal{E})$, a target node feature $x_i \in \mathbb{R}^C$, a source node feature $x_j \in \mathbb{R}^C$, and an edge feature $e_{ij} \in \mathbb{R}^E$, the aggregation operation integrates the features of neighboring nodes $\mathcal{N}(x_i)$ using the edge features. The updated node feature x'_i is given by:

$$x'_i = h\Big(x_i, g\big(x_j, \mathcal{N}(x_i); W_{\text{aggregate}}\big); W_{\text{update}}\Big), \quad (2)$$

where $W_{\text{aggregate}}$ and W_{update} are the learnable weights of the aggregation and update operations, respectively and h is a nonlinear activation function. To im-

prove the expressiveness of the Max-Relative convolution layer, edge attributes are incorporated using a conditional gating mechanism. This is mathematically formulated as:

$$g(\cdot) = x_i'' = [x_i, \max\left[w_{ji} \cdot (\{x_j - x_i \mid x_j \in \mathcal{N}(x_i)\})\right]], w_{ji} = \alpha\big(W_{\text{Gat}} \cdot e_{ji}\big). \quad (3)$$

Here, $W_{\text{Gat}} \in \mathbb{R}^{C \times E}$ and α denotes a softmax or sigmoid function. The final update operation is given by:

$$h(\cdot) = x_i' = x_i'' W_{\text{update}} + b_h, \quad (4)$$

where b_h represents the bias term.

3 Experimental Setup

Dataset. This study utilizes the MSSEG-II challenge dataset, which includes 3D FLAIR scans from 100 MS patients across 15 MRI scanners [5]. The dataset consists of 40 training scans (11 scans without new lesions) and 60 test scans (28 were originally labeled as lesion-free). For the test cases, the challenge organizers updated the ground truth (GT), introducing new lesions in two previously lesion-free cases while also increasing the number and volume of new lesions in 32 cases. Thus, the revised test set includes 35 cases with new lesions and 25 without. This study uses the updated consensus GT.

Pre-processing. The MSSEG-II dataset's rigidly registered FLAIR longitudinal scans were preprocessed, including brain extraction via ROBEX, N4 bias field correction with ITK, and histogram matching for intensity normalization of the training set [16]. The Demons algorithm was applied to compute deformation fields between time points, capturing only lesion changes due to prior rigid registration [25].

Implementation Details. We trained the network using 3D patches of size 16×16×16 with a step size of 8×8×8, extracted from the FLAIR modality of the MSSEG-II challenge's training set (40 patient volumes). To address the class imbalance between lesion and non-lesion voxels, patches centered on each GT voxel were included. The patches were split into 80% for training and 20% for validation. The model was trained for 100 epochs with early stopping (patience of 5), using a batch size of 32 and a random seed of 42 on an NVIDIA A100 GPU with PyTorch Geometric. During inference, a sliding window of 16×16×16 patches is applied.

Evaluation. DEFUSE-MS was evaluated in two scenarios to thoroughly assess its performance. First, new lesion detection accuracy was analyzed using detection and segmentation metrics on 35 patients (out of 60) who exhibited at least one new lesion in the follow-up testing set. The evaluation metrics included lesion detection F1-score, sensitivity (SensL), precision (PPVL), and the segmentation dice score (DSC). Second, the specificity of the method was assessed using data from 25 patients with no new T2-w lesions by calculating the mean volume (in mm^3) of falsely predicted lesions. All metrics were computed

Table 1: Lesion detection and segmentation results on the MSSEG-II challenge test set: Comparison between our proposed method and alternative approaches. The results are reported as the mean ± standard deviation. Only from the model's output, small lesions, smaller than 3 mm^3, were excluded from F1-score computations.

Method	F1-score ↑	SensL ↑	PPVL ↑	DSC ↑	No.FPs ↓	Vol. mm^3 ↓
(w/ DF)						
DEFUSE-MS (DFLearned, w/ Spatial)	**0.65 ± 0.30**	**0.74 ± 0.28**	**0.65 ± 0.33**	**0.55 ± 0.24**	0.12 ± 0.33	1.50 ± 4.77
DEFUSE-MS (DFLearned, w/o Spatial)	0.64 ± 0.32	0.70 ± 0.31	0.63 ± 0.35	0.54 ± 0.25	0.28 ± 1.02	2.13 ± 8.62
DEFUSE-MS (DFMax, w/ Spatial)	0.56 ± 0.29	0.68 ± 0.29	0.57 ± 0.33	0.50 ± 0.26	0.88 ± 2.40	20.66 ± 79.89
DEFUSE-MS (DFMax, w/o Spatial)	0.55 ± 0.33	0.64 ± 0.33	0.57 ± 0.36	0.50 ± 0.27	0.20 ± 0.82	0.96 ± 4.12
SegFormer3D [18]	0.40 ± 0.36	0.49 ± 0.36	0.44 ± 0.41	0.36 ± 0.28	0.48 ± 1.66	5.30 ± 22.26
SlimUNETR [17]	0.45 ± 0.35	0.44 ± 0.36	0.55 ± 0.42	0.39 ± 0.31	0.44 ± 1.83	4.65 ± 19.44
UNeXt [27]	0.48 ± 0.35	0.57 ± 0.35	0.52 ± 0.39	0.46 ± 0.29	0.28 ± 0.84	2.60 ± 7.79
UNETER [11]	0.41 ± 0.33	0.57 ± 0.35	0.41 ± 0.36	0.41 ± 0.27	0.44 ± 0.96	3.32 ± 7.93
TransUNet [4]	0.42 ± 0.35	0.51 ± 0.37	0.44 ± 0.38	0.37 ± 0.28	1.60 ± 4.04	27.57 ± 83.75
UNet [19]	0.37 ± 0.35	0.38 ± 0.36	0.43 ± 0.39	0.32 ± 0.28	0.08 ± 0.40	0.27 ± 1.33
(w/o DF)						
DEFUSE-MS (NoDF, w/ Spatial)	**0.55 ± 0.34**	**0.63 ± 0.35**	**0.58 ± 0.37**	**0.49 ± 0.27**	0.20 ± 0.58	5.76 ± 18.07
DEFUSE-MS (NoDF, w/o Spatial)	**0.55 ± 0.35**	0.59 ± 0.35	**0.58 ± 0.37**	0.46 ± 0.28	**0.0 ± 0.0**	**0.0 ± 0.0**
SegFormer3D [18]	0.43 ± 0.36	0.49 ± 0.37	0.48 ± 0.40	0.40 ± 0.30	0.80 ± 3.40	14.19 ± 68.26
SlimUNETR [17]	0.48 ± 0.36	0.59 ± 0.36	0.51 ± 0.39	0.44 ± 0.31	1.32 ± 4.60	37.71 ± 157.65
UNeXt [27]	**0.55 ± 0.34**	0.59 ± 0.37	0.57 ± 0.37	0.46 ± 0.30	0.04 ± 0.20	0.15 ± 0.77
UNETER [11]	0.46 ± 0.33	0.65 ± 0.32	0.45 ± 0.36	0.42 ± 0.27	1.68 ± 4.71	39.44 ± 169.41
TransUNet [4]	0.45 ± 0.35	0.57 ± 0.36	0.45 ± 0.37	0.40 ± 0.27	1.44 ± 5.23	47.11 ± 211.99
UNet [19]	0.27 ± 0.36	0.29 ± 0.38	0.34 ± 0.41	0.25 ± 0.27	0.12 ± 0.60	1.14 ± 5.70
(SOTA: Results are reported on the outdated test set (60 cases: 32 with new lesions and 28 without) [7])						
mediaire-B [12]	0.54 ± 0.35	0.69 ± 0.39	0.49 ± 0.36	0.44 ± 0.30	0.54 ± 0.84	29.23 ± 58.31
MedICL [13]	0.50 ± 0.33	0.74 ± 0.37	0.45 ± 0.34	0.51 ± 0.29	0.54 ± 0.84	12.71 ± 39.68
VicorobCascade [21]	0.50 ± 0.37	0.53 ± 0.40	0.52 ± 0.39	0.42 ± 0.32	0.46 ± 1.90	11.56 ± 41.10
Empenn [15]	0.53 ± 0.32	0.59 ± 0.37	0.53 ± 0.33	0.42 ± 0.26	0.29 ± 0.46	4.26 ± 9.0
SNAC [3]	0.51 ± 0.35	0.66 ± 0.40	0.47 ± 0.35	0.48 ± 0.29	0.32 ± 0.94	5.73 ± 19.75
Expert 1	0.68 ± 0.34	0.61 ± 0.36	0.83 ± 0.35	0.61 ± 0.32	0.0 ± 0.0	0.0 ± 0.0
Expert 2	0.58 ± 0.36	0.53 ± 0.36	0.73 ± 0.39	0.54 ± 0.33	0.0 ± 0.0	0.0 ± 0.0
Expert 3	0.58 ± 0.35	0.50 ± 0.36	0.77 ± 0.40	0.55 ± 0.34	0.0 ± 0.0	0.0 ± 0.0
Expert 4	0.49 ± 0.37	0.41 ± 0.37	0.70 ± 0.45	0.44 ± 0.34	0.0 ± 0.0	0.0 ± 0.0

using the animaSegPerfAnalyzer from the Anima toolbox, as described in [6].

Ablation Studies. We conducted ablation studies to evaluate the contributions of different components. DEFUSE-MS (DFLearned), which utilizes learned DF embeddings as temporal edge attributes, was compared with two variants: DEFUSE-MS (DFMax), which applies max pooling to the DF without learning embeddings, and DEFUSE-MS (NoDF), which does not use any temporal edge attributes. These comparisons highlight the importance of learned DF embeddings in capturing temporal dependencies. Additionally, we evaluated the role of spatial edge attributes by testing the model with and without them.

4 Results and Discussion

For quantitative results, Table 1 shows the F1-score, SensL, PPVL, and DSC of the DEFUSE-MS and several baseline models [18,17,27,11,4,19]. DEFUSE-MS (DFLearned) significantly outperformed the other two variants (DFMax and NoDF) and all the other CNN-based and Transformer-based models ($p < 0.05$). Although many baselines lack temporal modeling, they are widely used in medical segmentation, highlighting DEFUSE-MS's strength in capturing longitudinal dynamics. For qualitative results, Fig. 2 presents a visual example of the

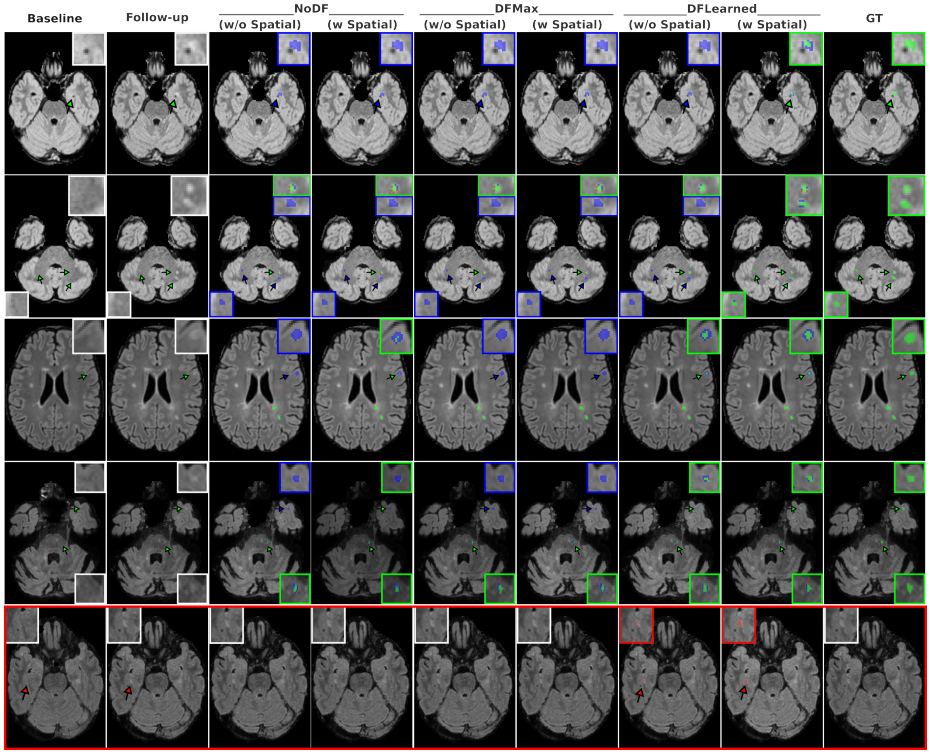

Fig. 2: Examples of new lesion detection. For the predicted segmentation masks, green, red, and blue represent true positives, false positives, and false negatives, respectively. The last row presents a case where only DFLearned detects false positives; however, we suspect the segmented region to be a new lesion, two of the human raters also classify this region as a new lesion.

DEFUSE-MS model's performance. Each column corresponds to the baseline image, follow-up image, segmentation results from the DEFUSE-MS approaches (NoDF, DFMax, DFLearned), and the GT mask.

Learned DF Embeddings. The impact of different DF representations on lesion detection performance was evaluated by comparing two variants of the DEFUSE-MS model: one using maxpooled DF (DFMax) and the other employing learned DF embeddings (DFLearned). The model with learned DF embeddings significantly outperformed the maxpooled variant, demonstrating superior F-score, SensL, PPVL and DSC in detecting new T2-w lesions ($p < 0.05$). This performance gap can be attributed to the information loss associated with maxpooling, which reduces spatial resolution and discards fine-grained deformation patterns crucial for accurate lesion localization. In contrast, learned embeddings adaptively capture complex spatial and temporal relationships, enhancing the model's ability to distinguish between lesions and normal anatomical variations.

Additionally, the learned embeddings provide richer contextual features that effectively support the spatiotemporal GNN layers in modeling temporal dependencies across follow-up scans. These findings emphasize the value of context-aware, learnable representations over static pooling operations, especially in detecting subtle pathological changes over time.

Spatial Edge Attributes. The impact of adding the Chebyshev distance as a spatial edge attribute in DEFUSE-MS varied depending on the temporal edge attribute used. In DEFUSE-MS (DFMax) and DEFUSE-MS (NoDF), which utilize maxpooled DF and no DF, respectively, the addition of the spatial edge attribute increased sensitivity (SensL) by enhancing spatial context and detecting subtle lesion boundaries. However, in follow-up cases without new lesions, both models showed an increase in the number and volume of false positives (FPs). This can be attributed to the over-sensitivity of the spatial edge attribute to anatomical variations, leading to the misclassification of normal structures as lesions. The lack of fine-grained temporal context in DFMax and the absence of temporal edge attributes in NoDF caused over-reliance on spatial discontinuities, contributing to FPs. In contrast, DEFUSE-MS (DFLearned), which uses learned DF embeddings, outperformed the other models across all metrics, including reducing FPs and their volumes in no-lesion cases. The learned embeddings effectively captured complex temporal deformations, enabling the model to distinguish between true pathological changes and normal anatomical variability. The synergy between the learned temporal context and the spatial edge attribute allowed the model to better generalize across patients, maintaining high sensitivity and precision. These results emphasize the value of using learnable temporal embeddings in combination with spatial edge attributes to balance sensitivity and specificity in lesion detection.

Model Architecture Comparison. Our study found that the integration of the DF into different model architectures yields varying results depending on the model type. In attention-based models, stacking the DF with baseline/follow-up images leads to decreased performance due to information overload, noise, and misalignment, which confuses the attention mechanism and hinders its ability to focus on relevant features. These models, optimized for learning contextual relationships within image features, struggle with the spatial transformations captured by the DF. In contrast, the UNet architecture, with its encoder-decoder structure and skip connections, significantly improves performance when stacking the DF with images ($p < 0.05$). The UNet effectively fuses spatial information from the DF with image features, leveraging its ability to maintain spatial consistency and learn local and global dependencies. Compared to UNet, DEFUSE-MS (W/ DF) may have more false positives due to over-segmentation, but it still achieves better overall detection metrics like Dice and F1. In contrast, DEFUSE-MS effectively integrates the DF as an edge attribute within spatiotemporal GNN layers, enhancing temporal dependency learning and spatial relationship modeling, thus benefiting from the DF. These findings suggest that targeted integration methods, such as using the DF as an edge attribute,

are more effective than direct stacking for leveraging spatial transformations in lesion detection tasks.

5 Conclusion

We proposed DEFUSE-MS, a novel approach for MS new lesion detection, which formulates lesion identification in 3D MRI as a heterogeneous spatiotemporal GNN within an encoder-decoder framework. By integrating learned DF embeddings as temporal edge attributes, DEFUSE-MS effectively captures spatial and temporal dependencies, improving sensitivity and precision. This approach outperforms traditional DF representations and addresses the challenges faced by attention-based models, which struggle with information overload and misalignment. DEFUSE-MS adapts to complex spatiotemporal changes, offering more accurate lesion localization and demonstrating the superiority of learned DF embeddings for automated MS new lesion detection in 3D MRI. Future work will enhance DEFUSE-MS by integrating more imaging modalities and improving generalizability across diverse populations.

Acknowledgments. This publication is funded by the UAE National Multiple Sclerosis Society (NMSS).

Disclosure of Interests. The authors have no competing interests in the paper as required by the publisher.

References

1. Altay, E.E., Fisher, E., Jones, S.E., Hara-Cleaver, C., Lee, J.C., Rudick, R.A.: Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. JAMA neurology **70**(3), 338–344 (2013)
2. Cabezas, M., Corral, J., Oliver, A., Díez, Y., Tintoré, M., Auger, C., Montalban, X., Lladó, X., Pareto, D., Rovira, À.: Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. American Journal of Neuroradiology **37**(10), 1816–1823 (2016)
3. Cabezas, M., Luo, Y., Kyle, K., Ly, L., Wang, C., Barnett, M.: Estimating lesion activity through feature similarity: A dual path Unet approach for the MSSEG2 MICCAI challenge. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure p. 107 (2021)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Commowick, O., Combès, B., Cervenansky, F., Dojat, M.: Editorial: Automatic methods for multiple sclerosis new lesions detection and segmentation. Frontiers in Neuroscience **17** (2023)
6. Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., Pop, S.C., Girard, P., Ameli, R., Ferré, J.C., et al.: Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Scientific reports **8**(1), 13650 (2018)

7. Commowick, O., Masson, A., Combes, B., Camarasu-Pop, S., Cervenansky, F., Kain, M., Lion, S., Casey, R., Dojat, M., Cotton, F.: MICCAI 2021 MSSEG-2 challenge quantitative results (Oct 2021), https://doi.org/10.5281/zenodo.5775523
8. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Gold, R., Kappos, L., Arnold, D.L., Bar-Or, A., Giovannoni, G., Selmaj, K., Tornatore, C., Sweetser, M.T., Yang, M., Sheikh, S.I., et al.: Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. New England Journal of Medicine **367**(12), 1098–1107 (2012)
10. Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E.: Vision GNN: An image is worth graph of nodes. Advances in neural information processing systems **35**, 8291–8303 (2022)
11. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
12. Hitziger, S., Ling, W.X., Fritz, T., D'Albis, T., Lemke, A., Grilo, J.: Triplanar U-Net with lesion-wise voting for the segmentation of new lesions on longitudinal MRI studies. Frontiers in Neuroscience **16**, 964250 (2022)
13. Kamraoui, R.A., Ta, V.T., Manjon, J.V., Coupé, P.: Image quality data augmentation for new MS lesion segmentation. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure **37** (2021)
14. Lladó, X., Ganiler, O., Oliver, A., Martí, R., Freixenet, J., Valls, L., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À.: Automated detection of multiple sclerosis lesions in serial brain MRI. Neuroradiology **54**, 787–807 (2012)
15. Masson, A., Le Bon, B., Kerbrat, A., Edan, G., Galassi, F., Combes, B.: A nnUnet implementation of new lesions segmentation from serial FLAIR images of MS patients. MSSEG-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure p. 5 (2021)
16. Nyúl, L.G., Udupa, J.K., Zhang, X.: New variants of a method of MRI scale standardization. IEEE transactions on medical imaging **19**(2), 143–150 (2000)
17. Pang, Y., Liang, J., Huang, T., Chen, H., Li, Y., Li, D., Huang, L., Wang, Q.: Slim UNETR: Scale hybrid transformers to efficient 3D medical image segmentation under limited computational resources. IEEE Transactions on Medical Imaging **43**(3), 994–1005 (2024)
18. Perera, S., Navard, P., Yilmaz, A.: SegFormer3D: an efficient transformer for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4981–4988 (2024)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
20. Salem, M., Cabezas, M., Valverde, S., Pareto, D., Oliver, A., Salvi, J., Rovira, À., Lladó, X.: A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. NeuroImage: Clinical **17**, 607–615 (2018)
21. Salem, M., Ryan, M.A., Oliver, A., Hussain, K.F., Lladó, X.: Improving the detection of new lesions in multiple sclerosis with a cascaded 3D fully convolutional neural network approach. Frontiers in Neuroscience **16**, 1007619 (2022)

22. Salem, M., Valverde, S., Cabezas, M., Pareto, D., Oliver, A., Salvi, J., Rovira, À., Lladó, X.: A fully convolutional neural network for new T2-w lesion detection in multiple sclerosis. NeuroImage: Clinical **25**, 102149 (2020)
23. Schmidt, P., Pongratz, V., Küster, P., Meier, D., Wuerfel, J., Lukas, C., Bellenberg, B., Zipp, F., Groppa, S., Sämann, P.G., et al.: Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging. NeuroImage: Clinical **23**, 101849 (2019)
24. Tahghighi, P., Zhang, Y., Souza, R., Komeili, A.: Enhancing new multiple sclerosis lesion segmentation via self-supervised pre-training and synthetic lesion integration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 263–272. Springer (2024)
25. Thirion, J.P.: Image matching as a diffusion process: an analogy with Maxwell's demons. Medical image analysis **2**(3), 243–260 (1998)
26. Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., Correale, J., Fazekas, F., Filippi, M., Freedman, M.S., et al.: Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. The Lancet Neurology **17**(2), 162–173 (2018)
27. Valanarasu, J.M.J., Patel, V.M.: UNeXt: MLP-based rapid medical image segmentation network. In: International conference on medical image computing and computer-assisted intervention. pp. 23–33. Springer (2022)
28. Wang, Y., Jiang, C., Luo, S., Dai, Y., Zhang, J.: Graph neural network enhanced dual-branch network for lesion segmentation in ultrasound images. Expert Systems with Applications **256**, 124835 (2024)
29. Wu, Y., Wu, Z., Shi, H., Picker, B., Chong, W., Cai, J.: Coactseg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 3–13. Springer (2023)