ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Multi-View Graph Neural Network for Image Segmentation : Intermediate vs Late Fusion

E. Karam[1,2,3]N. Jrad[2], P. Coupeau[1], D. Tobiano[2], J.-B. Fasquel[1], F. Abdallah[3,4] *

1- Université d'Angers - LARIS SFR MATHSTIC, F-49000 Angers - France

2- Université Catholique de l'Ouest - LARIS SFR MATHSTIC, F-49000 Angers - France

3- Lebanese University - Doctoral School of Science and Technology, Beirut - Liban

4- Université de Lorraine - Laboratoire LCOMS, Saint-Dié-des-Vosges - France

Abstract. Representing an image as a graph captures its spatial and contextual relationships effectively. Using Graph Neural Networks (GNNs) on graph-based images has considerably enhanced image segmentation. This paper investigates Multi-View GNNs for image segmentation, comparing Intermediate and Late Fusion methods. Experiments show that Intermediate Fusion achieves high accuracy on synthetic data by integrating relational features upfront. On a real dataset, Late Fusion methods, particularly RVCons, outperform Intermediate Fusion by dynamically aggregating multi-view predictions. Indeed, Late Fusion effectively mitigates issues arising from view-specific noise and variance. The results underscore the complementary strengths of both fusion strategies.

1 Introduction

Image segmentation is crucial for tasks like autonomous driving and medical imaging, aiding analysis and understanding spatial relationships between objects [1]. Deep learning has significantly improved segmentation accuracy, but often struggled to capture complex spatial and contextual dependencies. High-level structural information can be effectively represented using graphs, where nodes embody objects or regions of interest, and edges encode spatial relationships or contextual information between these elements. Traditional graph matching methods, typically framed as quadratic assignment problems, can represent spatial relations accurately but are computationally demanding [2]. In contrast, combining Deep Neural Networks (DNNs) with Graph Neural Networks (GNNs) offers a powerful alternative to explore graph-based structures. Here, graphs model the structural information, while GNNs enable learning and decision-making from these graph representations, enhancing applications across domains such as molecular modeling and image segmentation [3, 4]. Besides, incorporating heterogeneous graphs or multi-view representations can enhance image processing tasks by capturing various complementary aspects of the data, such as texture, color, and spatial relationships, which are crucial for accurate segmentation. Multi-view GNNs (MVGNNs) have recently demonstrated the

*The authors would like to express their gratitude to Angers Loire Métropole and Université Catholique de l'Ouest for their financial support and contributions to this project.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

capability to manage these heterogeneous relationships by integrating diverse data perspectives. By representing different aspects of the data, MVGNNs facilitate the integration of complementary information, and improve classification accuracy [5]. Research studies illustrate the advantages of fusing multiple graph views to improve classification accuracy in complex datasets [6].
This paper proposes two MVGNN architectures for segmentation: an Intermediate Fusion (IF) model that concatenates the resulting GNN outputs in an end-to-end framework, and a Late Fusion (LF) model that processes each view independently and aggregates predictions for enhanced robustness. We compare these architectures across synthetic and real-world data sets, demonstrating the strengths of each fusion approach under varied conditions.

2 Methods Overview

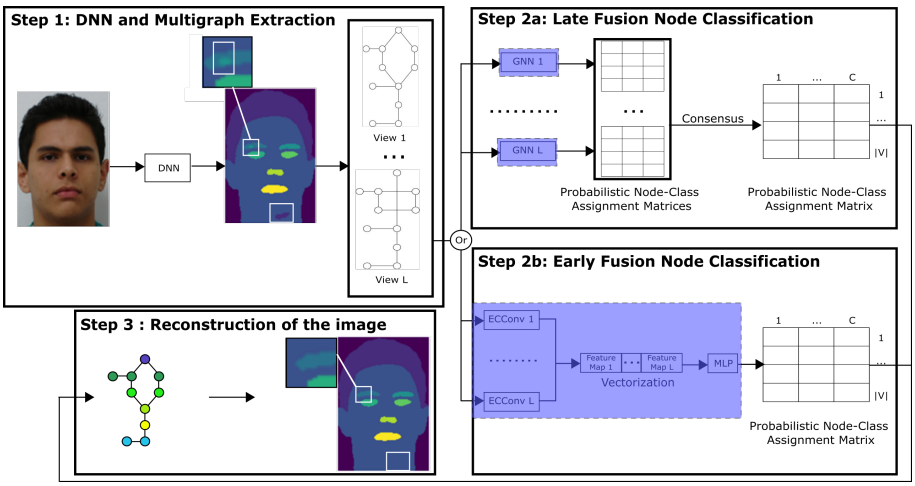

Fig. 1: Overview of the proposed approach

2.1 From Images to Multigraphs and GNN Processing (Step 1)

Given an image to be segmented, a DNN generates a segmentation map $S \in \mathbb{R}^{P \times C}$, where P represents the spatial dimensions of the image and C the number of classes, associating each pixel p with probabilities $S(p, c)$ of belonging to each class c. Based on these probabilities, a set R is constructed, consisting of all connected components of pixels that preliminarily belong to the same class. These regions are used to form a graph $G = (V, E, X, A)$, where V denotes the nodes (regions), E the edges between nodes, X the node attributes, and A the edge attributes of dimension d. To capture diverse spatial relationships, multiple graph views $G_1, \ldots, G_L$ are created, each representing specific spatial or structural attributes (e.g. distance, orientation), thus forming a multigraph

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

$M = (G_1, \ldots, G_L)$ with L distinct views. After Step 1, the process branches into Step 2a or Step 2b.

2.2 Late Fusion - General Consensus Function (Step 2a)

For LF method, each graph view G_l is independently processed by a GNN followed by a single layer perceptron. The GNN used includes a convolution layer, implemented as an edge-conditioned convolution operator (ECConv) [4], to locally aggregate information by updating node attributes. The SLP computes class probabilities based on the updated node features. The learning process is performed step by step.

To unify the information from each view, we define a consensus function that operates on the probability membership outputs of the GNNs across the views. Let $P_l \in \mathbb{R}^{|V| \times C}$ represent the probability membership matrix predicted by the GNN for view G_l. Each entry $P_l(i, j)$ in the matrix denotes the probability that node i belongs to class j in view G_l. The consensus function, denoted as Consensus$(P_1, P_2, \ldots, P_L)$, aggregates these probability membership matrices from all views to produce a unified segmentation output $P_{\text{consensus}} \in \mathbb{R}^{|V| \times C}$. This function can take various forms adapting dynamically to the characteristics of each view and leveraging the strengths of the combined predictions.

2.3 Intermediate Fusion - General Concatenation Function (Step 2b)

For IF, each view is processed through a dedicated GNN convolutional layer, capturing its unique features, and the resulting representations are concatenated to form a comprehensive feature space. This concatenated representation effectively integrates the information from all views early in the model's workflow and is then passed through a multi-layer perceptron (MLP) for final predictions. The entire process is trained end-to-end in a single optimization step, ensuring that the features from all views are jointly refined and leveraged throughout the model early on.

Step 3 reassigns for each region its class based on the predictions of Step 2a or 2b, to reconstruct the final segmentated image.

3 Experimental setup and Datasets

3.1 Datasets

A synthetic dataset is created based on a reference image containing 7 regions/classes: 6 distinct classes and a background class (Fig. 2). Each region, excluding the background, is represented as a node with a class probability vector. 3 graph views capture different spatial relations ; (1) undirected edges with a distance attribute, (2) directed edges indicating vertical relationships, and (3) directed edges indicating horizontal relationships.
A real dataset, FASSEG[1], is also used. It contains 70 human face images with

[1]FASSEG: https://github.com/massimomauro/FASSEG-repository

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

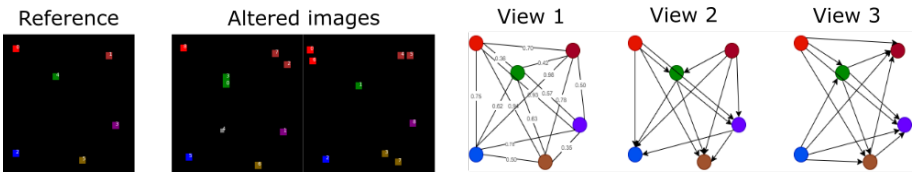

Fig. 2: Overview of the synthetic dataset: Reference image, its altered versions, and corresponding multi-view graph representations

9 classes (regions annotated as hair, eyes, etc and including background). 5 graph views are created based on spatial relationships. The first view represents distance-based min-max undirected edges. The second view encodes angular relationships as directed edges with three attributes. The third view captures boundary-sharing relationships through undirected edges, with shared space as the edge attribute. The fourth view represents vertical binary relationships as directed edges, while the fifth view represents horizontal binary relationships as directed edges.

3.2 Experimental Protocol

For the synthetic dataset, a set of 100 images for training/validation and 50 images for testing is generated. Subsets are created by randomly selecting images from the initial set as follows : 10 draws of 50, 20, and 10 training images, with respectively 15, 10, and 5 validation images; and 25 draws of 5 training and 5 validation images. A model is trained on each subset, with 300 epochs using validation accuracy as stopping criteria. An Adam optimizer (learning rate of 0.01, weight decay of 5e-4) and StepLR scheduler (decay factor 0.5 every 50 steps) are used.

For FASSEG, the dataset is split into 20 training/validation images and 50 test images, with U-Net as the segmentation backbone [7]. Subsets are created with 4 draws of 15 training/5 validation images and 4 draws of 5 training/2 validation images. Training parameters (epochs, stopping criteria, optimize and scheduler) are identical to those used for synthetic dataset. To evaluate accuracy, we use the Dice Similarity Coefficient (DSC), bounding box Dice index (B-DSC), and Hausdorff distance (HD) to measure overlap, spatial extent, and alignment.

For LF, a consensus function is used to obtain a consensus segmentation output. Consensus techniques used in our experiments include: Mean Consensus (averages class probabilities), Median Consensus (minimizes the influence of outliers), Majority Voting [8], RV Cons [9] (weights views using a similarity metric, RV, to refine consensus by reducing inconsistent views), and Attention-Based Consensus (dynamically weights views based on relevance, projecting them into a shared space to compute normalized attention scores for adaptive weighting).

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

4 Results

4.1 Synthetic Data

As seen in Table 1, consensus methods consistently outperform single-view models by effectively aggregating information across views. Concatenation stands out on synthetic data, maintaining high accuracy across all training sizes. The resilience of the consensus and concatenation methods suggests that these approaches provide robustness under limited data conditions.

Method	Number of Training Images			
	50 Images	**20 Images**	**10 Images**	**5 Images**
View 1	0.81 ± 0.02	0.74 ± 0.04	0.74 ± 0.03	0.70 ± 0.04
View 2	0.73 ± 0.04	0.67 ± 0.07	0.66 ± 0.06	0.63 ± 0.08
View 3	0.83 ± 0.02	0.76 ± 0.09	0.69 ± 0.05	0.68 ± 0.08
Mean Consensus	0.94 ± 0.02	0.94 ± 0.02	0.92 ± 0.03	0.87 ± 0.04
Median Consensus	0.91 ± 0.02	0.89 ± 0.02	0.87 ± 0.03	0.83 ± 0.05
Majority Voting	0.85 ± 0.02	0.82 ± 0.02	0.83 ± 0.04	0.78 ± 0.05
RVCons	0.90 ± 0.02	0.90 ± 0.02	0.88 ± 0.03	0.85 ± 0.04
Attention Consensus	0.93 ± 0.02	0.94 ± 0.02	0.92 ± 0.03	0.87 ± 0.04
Concatenation	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.99 \pm 0.00}$	$\mathbf{0.98 \pm 0.02}$	$\mathbf{0.96 \pm 0.01}$

Table 1: Mean and Standard Deviation accuracy of the synthetic data.

4.2 FASSEG Dataset

As shown in Table 2, single-view methods perform adequately with 15 training images, but degrade with 5. With 15 training images, the single-view model (View 1) performs similarly to multi-view models, while being simpler and less computationally expensive. In contrast, for 5 training images, multi-view methods, especially LF methods - RVCons, outperform single views by dynamically aggregating information across views. These findings underscore the adaptability of consensus approaches in handling real-world segmentation tasks.

Method	15 Training Images			5 Training Images		
	DSC	BDSC	HD	DSC	BDSC	HD
CNN	0.70 ± 0.01	0.59 ± 0.01	37.40 ± 1.11	0.71 ± 0.04	0.56 ± 0.06	54.13 ± 12.69
View 1	$\mathbf{0.70 \pm 0.01}$	$\mathbf{0.61 \pm 0.01}$	$\mathbf{27.05 \pm 0.98}$	0.68 ± 0.04	0.57 ± 0.04	53.34 ± 12.58
View 2	0.67 ± 0.04	0.57 ± 0.03	42.53 ± 18.57	0.62 ± 0.09	0.52 ± 0.09	74.34 ± 21.99
View 3	0.70 ± 0.01	0.61 ± 0.01	28.39 ± 0.40	0.67 ± 0.04	0.56 ± 0.05	59.86 ± 11.34
View 4	0.70 ± 0.01	0.61 ± 0.01	28.47 ± 1.19	0.50 ± 0.13	0.42 ± 0.12	119.9 ± 48.70
View 5	0.67 ± 0.05	0.58 ± 0.05	39.56 ± 18.04	0.58 ± 0.11	0.48 ± 0.10	92.73 ± 35.36
Mean Cons.	0.70 ± 0.01	0.61 ± 0.01	27.52 ± 0.97	0.71 ± 0.04	0.60 ± 0.05	41.95 ± 10.64
Median	0.70 ± 0.01	0.61 ± 0.01	27.65 ± 0.55	0.71 ± 0.04	0.60 ± 0.05	41.75 ± 9.43
Majority	0.70 ± 0.01	0.60 ± 0.01	27.88 ± 0.85	0.71 ± 0.04	0.60 ± 0.05	42.39 ± 10.41
RVCons	$\mathbf{0.70 \pm 0.01}$	$\mathbf{0.61 \pm 0.01}$	$\mathbf{27.48 \pm 0.79}$	$\mathbf{0.72 \pm 0.04}$	$\mathbf{0.60 \pm 0.05}$	$\mathbf{40.50 \pm 9.46}$
Attention	0.70 ± 0.01	0.61 ± 0.01	27.42 ± 1.07	0.71 ± 0.04	0.60 ± 0.05	41.33 ± 10.59
Concat.	0.70 ± 0.01	0.60 ± 0.01	29.31 ± 0.36	0.69 ± 0.04	0.58 ± 0.06	47.92 ± 10.92

Table 2: Performance metrics for different methods with DSC, BDSC, and HD for 15 and 5 Training Images.

Figure 3 provides a visual example of how RVCons improves segmentation by correcting errors found in individual views. Each single-view segmentation shows inconsistent errors caused by limited data. However, RVCons effectively integrates information across views, correcting errors and yielding a result that closely resembles the ground truth.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

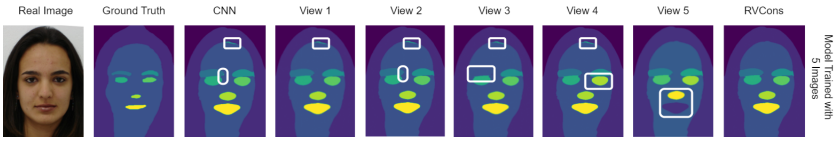

Fig. 3: Segmentation Example for image trained with 5 images with Error Regions Marked in White Boxes

5 Conclusion

This study evaluated the effectiveness of MVGNNs for image segmentation. By leveraging information from multiple views, MVGNN outperforms single-view GNNs, particularly under limited training data. Comparing IF and LF approaches, the results highlight the complementary strengths of both fusion strategies. IF excels in scenarios where intermediate integration of features helps uncover joint patterns across views, making it ideal for homogeneous datasets. In contrast, LF demonstrates its advantage in complex, noisy environments by preserving view-specific processing and enabling adaptive integration of predictions. These findings suggest that the choice of fusion strategy should be tailored to the nature of the dataset. This comparison underscores the importance of understanding the interplay between data characteristics and fusion strategies when designing MVGNN architectures for multi-view image segmentation tasks.

References

[1] L. Waikhom and R. Patgiri. Graph neural networks: Methods, applications, and opportunities. *CoRR*, abs/2108.10733, 2021. Available: `https://arxiv.org/abs/2108.10733`.

[2] J. Chopin, J.-B. Fasquel, H. Mouchère, R. Dahyot, and I. Bloch. Model-based inexact graph matching on top of dnns for semantic scene understanding. *Computer Vision and Image Understanding*, 235:103744, 2023.

[3] S. Chen, A. Wulamu, Q. Zou, H. Zheng, L. Wen, X. Guo, H. Chen, T. Zhang, and Y. Zhang. Md-gnn: A mechanism-data-driven graph neural network for molecular properties prediction and new material discovery. *Journal of Molecular Graphics and Modelling*, 123, 2023.

[4] P. Coupeau, J.-B. Fasquel, and M. Dinomais. On the use of gnn-based structural information to improve cnn-based semantic image segmentation. *Journal of Visual Communication and Image Representation*, 101:104167, 2024.

[5] J. Wang, L. Wu, H. Zhao, and N. Jia. Multi-view enhanced zero-shot node classification. *Information Processing and Management*, 60(6):103479, 2023.

[6] E. Karam, N. Jrad, P. Coupeau, J.-B. Fasquel, and F. Abdallah. Multi-view graph neural network for semantic image segmentation. In *International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2024.

[7] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241, 2015.

[8] L. Lam and S. Y. Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(5):553–568, 1997.

[9] N. Niang and M. Ouattara. Weighted multiblock clustering. In *Book of Abstracts*, page 135.