Topology-Aware Vision Transformers for Enhanced Scene Recognition

Yunxi Wang[1], Shuaiyu Liu[1], Qiling Li[2], Yazhou Ren[1,3*], Xiaorong Pu[1,3]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2] School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan, China
[3] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China
{2023080910017, 202422081319}@std.uestc.edu.cn, gushujia@cdu.edu.cn {yazhou.ren, puxiaor}@uestc.edu.cn

Abstract

Scene recognition (SR) is a fundamental task in computer vision (CV). In recent years, Transformer-based methods have achieved remarkable success in scene recognition tasks. Most existing approaches primarily rely on visual features, while failing to effectively model the structural relationships within scenes, which are crucial for accurate scene recognition. To this end, we propose **T**opology **A**ttention **N**etwork for **S**cene **R**ecognition (**TANSR**), an innovative method that leverages topological relationships from graphs to guide scene recognition. Specifically, **G**raph **A**ttention **M**ask **G**eneration **N**etwork (**GAMGN**) generates topology-aware masks from graph representations constructed by **G**raph **G**eneration **M**odule (**GGM**) and integrates them with patch embeddings by **T**opology **A**ttention **G**uidance (**TAG**), enabling the transformer's attention mechanism to incorporate topological information. Furthermore, we introduce an innovative attention-driven multimodal fusion strategy that integrates graph-derived topological cues with visual patch embeddings, substantially enhancing the transformer's capability to capture topological information and improving performance in complex scene recognition tasks. We evaluate TANSR on the benchmarks MIT-67, Scene-15 and SUN397, where it achieves consistent state-of-the-art (SOTA) performance, including **98.58%** accuracy on MIT-67.

Code — https://github.com/CyanCQC/TANSR

Introduction

Scene recognition, a fundamental task in computer vision, is crucial for various applications such as autonomous driving, human-computer interaction (HCI), virtual reality (VR), and augmented reality (AR). Early approaches primarily relied on global attribute descriptors and manual feature extraction methods (Xie et al. 2020), such as SIFT (Lowe 2004) and HOG (Dalal and Triggs 2005), to model visual properties. However, these methods achieved limited performance due to their shallow representation capability and limited capacity to effectively represent complex scene structures.

The advent of deep learning has revolutionized scene recognition, with CNN-based methods like DAG-CNNs (Yang and Ramanan 2015) becoming widely adopted for image representations. To address classification challenges in complex scenarios, the Vision Transformer (ViT), proposed by Dosovitskiy (2020), employs a multi-head self-attention mechanism to capture global image features, demonstrating excellent scalability and transferability. Building on this advancement, Niu, Ma, and Li (2024) introduced SC-ViT, a ViT architecture specifically tailored for scene recognition. Touvron et al. (2021) developed DeiT-B, while Said et al. (2023) proposed the Dual Multiscale Attention ViT, both further optimizing scene recognition and achieving SOTA results.

However, existing Transformer-based methods mainly focus on visual features, while neglecting the relationships between elements. As a result, they often fail to distinguish visually similar scenes with different spatial arrangements. Chen et al. (2020) introduced graph feature learning into

(a) Waiting room (b) Living room

Figure 1: Visually similar examples. Top: SLIC (Achanta et al. 2012) superpixel boundaries. Bottom: corresponding superpixel graphs, revealing distinct topological structures.

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

convolutional neural networks, effectively enhancing scene recognition performance and highlighting the critical role of structural relationships in this task. For instance, sidewalks and roadways are typically adjacent, shelves contain goods, and seats in a theater are always arranged in a surrounding relationship with the stage. These are essentially reflections of the topological relationships among different elements. Therefore, we collectively refer to them as topological features in this paper. Utilizing topological information helps to better distinguish scenes. As illustrated in Figure 1, scenes like living rooms and waiting rooms share similar elements—sofas, chairs, and tables—making them difficult to distinguish by appearance alone. However, their topological layouts differ markedly: living rooms typically feature compact furniture arrangements centered around a table, whereas waiting rooms exhibit more dispersed seating organized into smaller, independent groups.

A graph is composed of a node set and an edge set, which effectively represent topological structures by describing the properties of nodes and the connectivity between them. Combining scene images with their corresponding graph data for scene recognition enables the utilization of both spatial and topological structures of the scenes. Building upon these insights, we propose **T**opology **A**ttention **N**etwork for **S**cene **R**ecognition (TANSR), which integrates graph-based topological features into transformer networks. To the best of our knowledge, TANSR is the first transformer framework explicitly guided by graph-based topological features for scene recognition, bridging the gap between visual appearance and structural relationships. Specifically, we design a unified framework that explicitly embeds topological priors into transformer attention. The Graph Generation Module (GGM) first abstracts scene images into compact, semantically meaningful graphs via superpixel-based decomposition, enabling a structured representation of topological relationships. Building on this, the Graph Attention Mask Generation Network (GAMGN) learns topology-aware attention masks that dynamically align with transformer patches through a tailored graph attention mechanism (Velickovic et al. 2017). These masks are further fused by the proposed Topology Attention Guidance (TAG) module, which adaptively directs the transformer's attention towards structurally discriminative regions of the scene, allowing it to focus on topological relationships rather than solely relying on appearance-based cues. The major contributions of our study can be summarized as follows:

- We propose the GGM and GAMGN modules, which first transform images into feature-rich graphs and then utilize graph representations to encode topological relationships as attention mask vectors, enhancing the model's ability to capture structural information.
- We introduce the TAG method, which effectively leverages implicit structural relationships to improve scene recognition performance, presenting a novel attention-based multimodal fusion approach that strengthens ViT's capability to integrate spatial and topological cues.
- Our model achieves SOTA performance on MIT-67, Scene-15, and SUN397, demonstrating significant improvements, particularly in visually similar yet structurally distinct scenes where ViTs often struggle.

Related Work

Scene Recognition

Scene recognition (SR), a fundamental task in computer vision, supports a wide range of applications such as mobile imaging and autonomous driving. Traditional methods based on handcrafted features have struggled to cope with increasing scene complexity, leading to the emergence of deep learning approaches. For example, Liu et al. (2018) introduced a dictionary learning layer to enhance sparse scene representations; Lin et al. (2022) proposed a comprehensive representation to encode contextual object information; Niu, Ma, and Li (2024) developed SC-ViT to jointly exploit geometric details and channel contributions; and Said et al. (2023) designed dual multiscale attention to capture features at different scales. However, these methods primarily focus on visual appearance while overlooking the spatial organization and topological features of scene elements. As a result, their performance may degrade in scenarios where visual similarity masks essential structural differences, limiting their performance across diverse scenes.

In real-world scenes, similar objects and backgrounds frequently appear across different categories, making structural differences crucial for scene discrimination. Chen et al. (2020) introduced a layout graph network that significantly improved performance. Inspired by this, we propose a topology-aware attention mechanism that explicitly models spatial layouts to capture structural variability across scenes. By incorporating topological information, our model better distinguishes scenes with similar visual content, achieving more robust and reliable recognition.

Graph Generation from Images

Graph generation from images is a fundamental task in image analysis, widely applied to segmentation, scene recognition, and object detection. Existing methods fall into three types: pixel-based, region-based, and feature-based. Pixel-based approaches, such as GBT (Suzuki, Ueda, and Sklansky 1993) and SSC-MSF (Tarabalka et al. 2010), define pixel-level relationships, offering the finest granularity but at high computational cost. Feature-based methods, e.g., LOS-SOD (Lu, Mahadevan, and Vasconcelos 2014) and 2S-AGCN (Shi et al. 2019), construct graphs from deep features, capturing rich semantics while losing spatial details. Region-based methods generate nodes from image segments, reducing complexity but relying heavily on segmentation quality. Superpixel-based approaches, including simple linear iterative clustering (SLIC) (Achanta et al. 2012), SEEDS (Hsu and Ding 2013), and MaskSLIC (Irving 2016), are efficient representatives of this category.

In scene recognition, images contain rich semantics and clear structural layouts that permit reliable region decomposition. We adopt SLIC for graph generation because it efficiently produces boundary-preserving, spatially coherent superpixels, enabling stable and semantically consistent graph structures for topology-aware learning in TANSR.

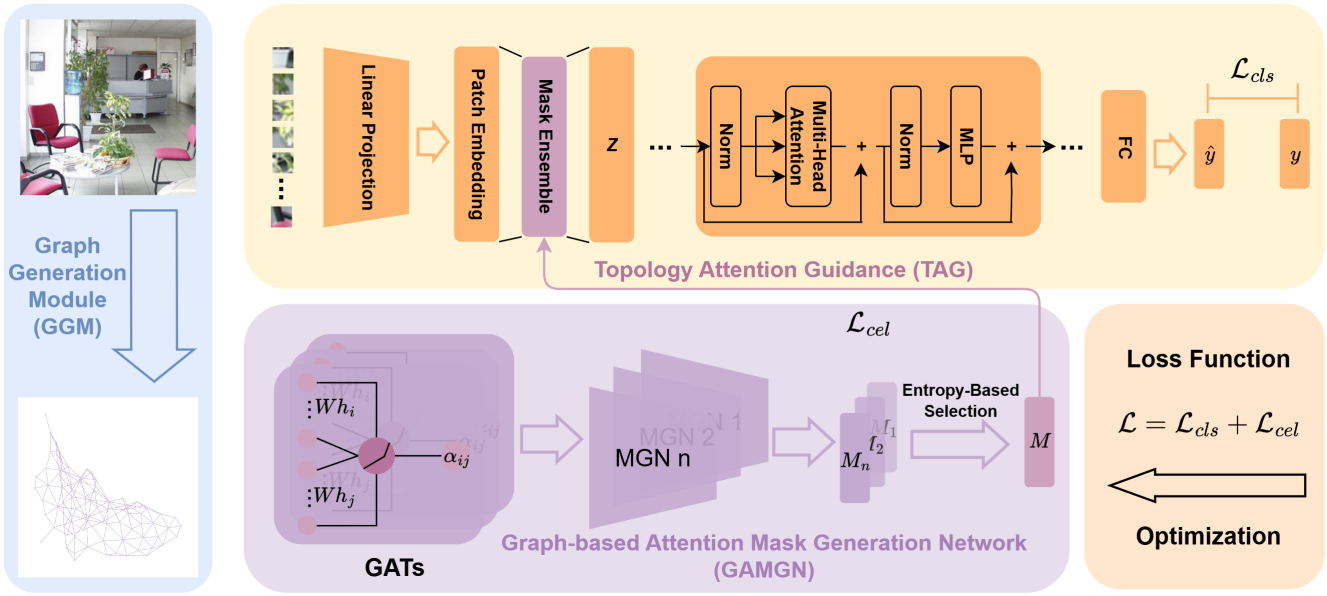


Figure 2: Model Overview. TANSR employs GGM to generate graphs from images, which are processed by GAMGN to extract features and produce label-specific masks. The selected masks are utilized in TAG to refine patch embeddings, allowing topology-aware features to guide ViT's attention.

Preliminaries

Superpixel-based Graph Construction Considering both computational complexity and the characteristics of the generated graph, we adopt the SLIC algorithm to segment images into perceptually coherent superpixels. SLIC applies k-means clustering in a five-dimensional feature space combining color (CIELAB components (l, a, b)) and spatial coordinates $(x_{\text{pos}}, y_{\text{pos}})$, producing compact and visually meaningful regions. Each superpixel represents a homogeneous region, and adjacency between superpixels reflects potential structural relationships.

Given a superpixel s_n, its basic feature is defined as the average of the pixel features:

$$\phi(s_n) = \frac{1}{|s_n|} \sum_{u \in s_n} F(u), \tag{1}$$

where $F(u)$ denotes the pixel feature (e.g., RGB values).

In addition, we consider the RGB mean and standard deviation within each superpixel region:

$$C_{\text{mean}} = \frac{1}{N} \sum_{i=1}^{N} C_i, \tag{2}$$

$$C_{\text{std}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (C_i - C_{\text{mean}})^2}, \tag{3}$$

where $C \in \{R, G, B\}$ and N is the number of pixels in the region. Each superpixel forms a graph node with attributes:

$$[x_{\text{pos}}, y_{\text{pos}}, R_{\text{mean}}, G_{\text{mean}}, B_{\text{mean}}, R_{\text{std}}, G_{\text{std}}, B_{\text{std}}],$$

and edges are defined by spatial adjacency.

Graph Attention Network (GAT) We adopt the Graph Attention Network (GAT) (Velickovic et al. 2017) to extract topological features. A GAT layer updates node v_i as:

$$h_{v_i}^{(l+1)} = \sigma \sum_{j \in \mathcal{N}(i)} \alpha_{ij} W^{(l)} h_{v_j}^{(l)}, \tag{4}$$

where α_{ij} is the normalized attention coefficient:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^\top [W^{(l)} h_{v_i}^{(l)} \| W^{(l)} h_{v_j}^{(l)}]\right)\right)}{\sum_{r \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(a^\top [W^{(l)} h_{v_i}^{(l)} \| W^{(l)} h_{v_r}^{(l)}]\right)\right)}. \tag{5}$$

We also adopt multi-head attention to stabilize learning:

$$h_{v_i}^{(l+1)} = \Big\|_{k=1}^{K_a} \sigma \sum_{v_j \in \mathcal{N}(v_i)} \alpha_{ij}^{(k)} W_k^{(l)} h_{v_j}^{(l)}, \tag{6}$$

where K_a is the number of attention heads, and $\|$ denotes concatenation along the feature dimension.

ViT Patch Embedding A ViT divides an image into P patches and projects them into an embedding matrix:

$$PE \in \mathbb{R}^{B \times P \times D}, \tag{7}$$

where B is the batch size and P is the patch number, and D the embedding dimension. These patch embeddings are then processed by multiple transformer layers with positional encodings and multi-head self-attention mechanisms.

Methodology

Overall Framework

TANSR integrates structural information into a transformer-based framework for topology-aware scene recognition. As illustrated in Figure 2, the pipeline consists of:

1. **Graph Generation Module (GGM)** to abstract superpixel-based structural graphs;
2. **Graph-based Attention Mask Generation Network (GAMGN)** to learn label-specific topology-aware masks aligned with transformer patches;
3. **Topology Attention Guidance (TAG)** to fuse the generated masks with patch embeddings, enabling topology-aware attention that enhances the model's performance.

The final class predictions are obtained via a fully connected classifier on the refined transformer output.

Graph Generation Module (GGM)

Based on SLIC segmentation, we construct a superpixel graph $G = (V, E)$, where each node represents a superpixel with attributes $[x_{\text{pos}}, y_{\text{pos}}, R_{\text{mean}}, G_{\text{mean}}, B_{\text{mean}}, R_{\text{std}}, G_{\text{std}}, B_{\text{std}}]$. Edges connect spatially adjacent superpixels.

After obtaining the initial superpixel graph from SLIC segmentation, we further refine the graph structure by assigning edge weights based on the color similarity between adjacent superpixels. Specifically, for each pair of connected nodes (v_i, v_j), we compute the Euclidean distance between their RGB mean vectors $\mu_i, \mu_j \in \mathbb{R}^3$ extracted from the corresponding superpixel regions:

$$d_{ij} = \|\mu_i - \mu_j\|_2. \tag{8}$$

The edge weight w_{ij} is defined as the inverse of distance:

$$w_{ij} = \frac{1}{d_{ij} + \epsilon}, \tag{9}$$

where $\epsilon = 10^{-6}$ is a small constant to prevent instability.

This design enforces stronger connectivity between visually similar superpixels while reducing the influence of structurally dissimilar regions. To ensure graph symmetry, bidirectional edges with identical weights are added, resulting in a refined adjacency matrix that better preserves local appearance consistency within the scene.

Graph Attention Mask Generation Network (GAMGN)

We apply an L-layer GAT to extract node embeddings $h_v^{(L)}$, which are globally pooled into a graph-level embedding:

$$g = \frac{1}{|V|} \sum_{v \in V} h_v^{(L)}, \tag{10}$$

This embedding is then divided into C class-specific subvectors g_c, each generating a label-specific mask via:

$$m_c = \text{Sigmoid}(W_c g_c). \tag{11}$$

where W_c is a class-specific linear projection matrix that maps g_c into the mask space, enabling each head to learn an independent topology-aware mask.

During training, for a sample with label y, the corresponding mask m_y is selected as the final mask:

$$m_{b,p}^{(f)} = m_y, \tag{12}$$

where b denotes the batch index and p the patch index.

During evaluation and testing, the mask with the highest entropy (as encouraged by our Contrastive Entropy Loss (CEL)) is selected as the final mask.

GAMGN leverages topological information within the scene structure through GAT layers and generates label-specific mask vectors via entropy-based optimization. These masks serve as the representation of topological information derived from the global graph features.

Topology Attention Guidance (TAG)

TAG integrates the mask $m^{(f)}$ with ViT patch embeddings:

$$z_{b,p,d} = PE_{b,p,d} \cdot m_{b,p}^{(f)}, \tag{13}$$

where topological information effectively guides ViT's attention by emphasizing structurally discriminative regions, thereby introducing topological information into the ViT through topology-aware attention mechanism.

Loss Function

We combine standard classification loss with a **Contrastive Entropy Loss (CEL)** that encourages high entropy for true-label masks to promote broader coverage of structurally relevant regions, while suppressing entropy for competing masks to enhance discriminative focus and reduce ambiguity. The loss function is formulated as follows:

$$\mathcal{L}_{cel} = \frac{1}{B} \sum_{b=1}^{B} \max\Big(0,\, \gamma - \big(H(m_{y_b}^{(b)}) - \min_{c \neq y_b} H(m_c^{(b)})\big)\Big), \tag{14}$$

where $H(\cdot)$ denotes the Shannon entropy. The hyperparameter γ is set to 1. The total loss is:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{cel}, \tag{15}$$

where the standard classification loss is:

$$\mathcal{L}_{cls} = -\frac{1}{B} \sum_{i=1}^{B} y_i \log(\hat{y}_i), \tag{16}$$

with y_i being the true label and $\hat{y}_i$ the predicted probability for the correct class. This combined loss ensures both accurate class predictions and robust mask generation, optimizing the model for complex scene recognition tasks.

Experiments

Datasets

To evaluate TANSR, we conducted experiments on three widely used baselines: MIT Indoor 67 (MIT-67) (Quattoni and Torralba 2009), Scene-15 (Lazebnik, Schmid, and Ponce 2006) and SUN397 (Xiao et al. 2010).

MIT-67 MIT-67 contains color images from 67 indoor scene categories and is a widely used benchmark for scene recognition. Its larger scale compared to Scene-15 makes it more suitable for our experimental analysis.

Scene-15 Scene-15 comprises grayscale images from 15 scene categories. We follow the standard protocol by randomly selecting 100 images per category for training and using the rest for testing, averaging results over ten random splits to evaluate performance on smaller datasets.

SUN397 SUN397 includes 397 categories and numerous color images covering natural and man-made scenes. SUN397 has large scale and high diversity, making it a more challenging benchmark for large-scale scene recognition.

Implementation Details

The ViT backbone is `vit_base_patch16_224` from the `timm` library (Wightman 2019), pretrained on ImageNet-1k with a patch size of 16. To adapt to smaller datasets, we applied data augmentation across all datasets, including random cropping, flipping, rotation, Gaussian blur, and Gaussian noise. Color jittering was applied only to MIT-67 and SUN397. Training was conducted on an NVIDIA RTX 4090 GPU and a Xeon Gold 6430 CPU.

The model was trained for 300 epochs with a batch size of 75 using the AdamW optimizer (Loshchilov 2017). The learning rate was set to `1e-5` for fine-tuning the ViT layers and `1e-3` for GAMGN, with a weight decay of `1e-3`. Images were resized to 224×224. The number of attention heads in the GAT layers was set equal to the number of categories in the dataset. To visualize embeddings in 2D, we used t-SNE (Van der Maaten and Hinton 2008), and Grad-CAM was employed to visualize attention maps.

Experimental Results

We selected three types of existing methods for comparison, including non-Transformer methods, hybrid methods, and Transformer-based methods. Among non-Transformer methods, we included DAG-CNN (Yang and Ramanan 2015), Mix-CNN (Hayat et al. 2016), Hybrid-CNNs (Xie et al. 2015), Multi-scale CNNs (Herranz, Jiang, and Li 2016), Dual CNN-DL (Liu et al. 2018), SDO (Cheng et al. 2018), MRNet (Lin et al. 2022), LGN (Chen et al. 2020) and EfficientNet-B7 (Tan and Le 2019). For hybrid methods, we included NEM (Saleknia et al. 2024). For Transformer-based methods, we included SC-ViT (Niu, Ma, and Li 2024), DeiT-B (Touvron et al. 2021), and DMS-ViT3 (Said et al. 2023). Although EfficientNet-B7 and DeiT-B were not originally designed for scene recognition, they have demonstrated strong performance as shown in (Said et al. 2023). Thus, we also take them into account. We use results from public code or reported experiments. Models without available code or unsupported on Scene-15 and SUN397 cannot be fairly re-run; their reported results are thus omitted.

On Scene-15, TANSR achieves an average accuracy of 97.12 ± 0.32 %, with results ranging from 96.65 % to 97.79 % across ten random train/test splits.

After comprehensive training and evaluation, TANSR consistently outperformed existing SOTA models across all benchmark datasets, as summarized in Table 1, which demonstrates TANSR's strong generalization capability and robustness in handling diverse scene recognition tasks.

Method	MIT-67 (%)	Scene-15 (%)	SUN397 (%)
Non-Transformer Methods			
DAG-CNN	77.50	92.90	56.20
Mix-CNN	79.63	–	57.47
Hybrid-CNNs	82.24	–	64.53
Multi-scale CNNs	86.04	95.18	–
Dual CNN-DL	86.43	96.03	70.13
SDO	86.76	95.90	73.41
MRNet	88.08	96.10	73.98
LGN	88.06	–	74.06
DPP-ResNeXt-101	90.82	–	<u>79.56</u>
EfficientNet-B7	95.60	<u>97.00</u>	–
Hybrid Methods			
NEM	89.30	96.80	–
Transformer-based Methods			
SC-ViT	90.60	–	77.79
DeiT-B	94.60	–	–
DMS-ViT3	<u>96.80</u>	–	–
TANSR (Ours)	**98.58**	**97.12**	**79.61**

Table 1: Comparison of **TANSR** with representative methods on **MIT-67**, **Scene-15**, and **SUN397**. The best results are boldfaced and the second results are underlined. TANSR consistently achieves SOTA performance.

Dataset	Nodes	Nodes Range	Edges	Avg Deg.
MIT-67	81.35±9.02	[19,104]	210.82±26.08	5.18
Scene-15	72.05±15.47	[3,100]	185.28±43.56	5.14
SUN397	72.33±12.24	[7,101]	185.16±34.56	5.12

Table 2: Graph statistics per patch (mean ± std) for different datasets generated by GGM.

Method	MIT-67 (%)	Δ
ViT-B	84.18	+0.00
ViT-B + GGM (CLS Concat)	84.85	+0.67
ViT-B + GGM + GAMGN (Concat)	92.99	+8.81
ViT-B + GGM + GAMGN + TAG (**TANSR**)	**98.58**	**+14.40**

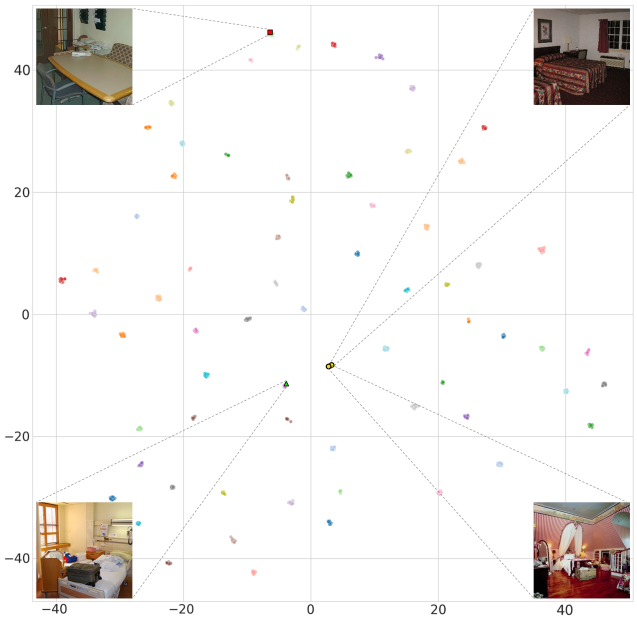
Table 3: Incremental improvements of TANSR components on the MIT-67 dataset. Results show the stepwise gains from GGM, GAMGN, and TAG over the ViT-B (the base ViT model `vit_base_patch16_224`).

Further Analysis

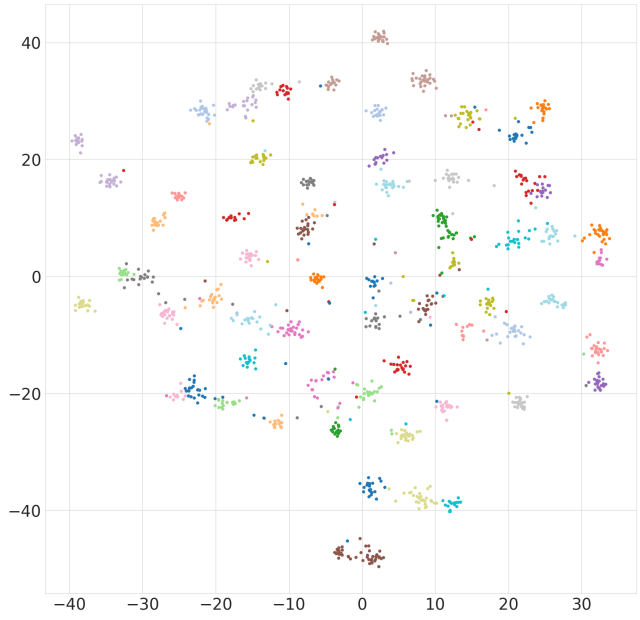
Generated Graph We analyze the statistical properties of GGM-generated graphs across datasets in terms of node and edge counts (mean ± std), node range, and average degree.

The denser graphs observed in MIT-67 align with the higher structural complexity of complex scenes (as its indoor layouts typically contain richer structural components). Meanwhile, the stable node degree ($\sim$ 5) across datasets confirms that GGM maintains consistent topology density, facilitating robust topology-aware learning in TANSR.

Ablation To evaluate the contribution of each component, we conducted a series of ablation studies. Removing all topology-aware modules (GGM, GAMGN, and TAG) re-



(a) TANSR



(b) ViT Only

Figure 3: Visualization of embeddings on MIT-67 dataset: Similar scene structures are positioned close to each other, while scenes with differing structures are spaced further apart. Clear separations between classes are observed.

duces the model to plain ViT, which shows weak class separability and a clear drop in accuracy (Table 3). Adding GGM introduces structural priors through a lightweight CLS-level concatenation, where the pooled graph embedding is fused with the CLS token before classification. Incorporating GAMGN further refines graph features and integrates them with visual representations via direct concatenation, yielding additional gains. Overall, these enhancements consistently improve recognition accuracy, and the t-SNE visualizations in Figure 3 show that TAG further sharpens category boundaries and produces more compact clusters.

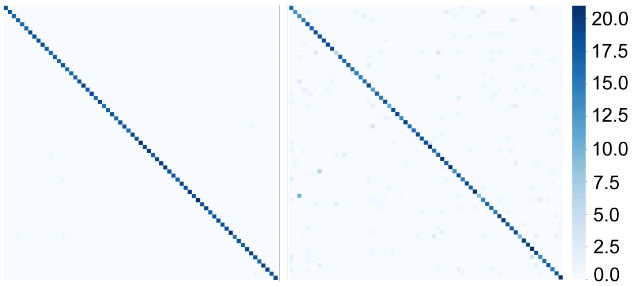


Figure 4: Confusion matrices of TANSR (left) and ViT-B (right) on MIT-67.

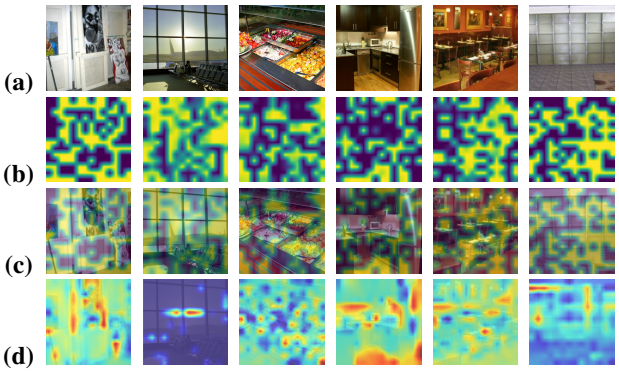


Figure 5: Visualization of samples: (a) original image; (b) generated mask; (c) blended mask; (d) mask-weighted attention map by Grad-CAM (Selvaraju et al. 2017).

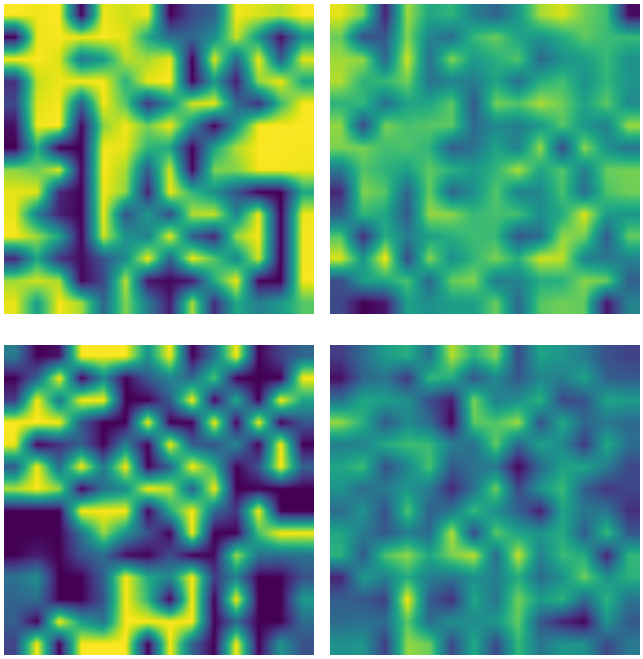
Figure 3a shows four sample scenes, including two bedrooms, one hospital room, and one meeting room. The bedroom images cluster closely, and the hospital room lies relatively near them due to structural similarity. In contrast, the meeting room, whose layout differs more substantially, appears farther away. These spatial relations show that TANSR effectively captures structural similarity across scenes. With the joint use of GGM, GAMGN, and TAG, TANSR enhances the ViT backbone's ability to encode structural and visual cues, producing more discriminative embeddings.

The confusion matrices in Figure 4 highlight that TANSR yields clearer inter-class distinctions and fewer misclassifications than the baseline ViT.

The visualizations of samples are shown in Figure 5, which include their TAG masks, blended heatmaps, and mask-weighted attention heatmaps generated by Grad-CAM. These visualizations demonstrate that the model effectively identifies and utilizes the topological features present in scene images. As a result, the model is able to capture the topological structures within the image, emphasizing the structural relationships between patches and improving scene recognition performance.

Dataset	Learning Rate				Batch Size			
	1e-1	1e-2	**1e-3**	1e-4	**75**	100	125	150
Scene-15	96.11	96.98	**97.12**	96.21	**97.12**	97.10	96.99	97.03
MIT-67	92.71	98.18	**98.58**	97.90	**98.58**	98.32	97.48	98.46

Table 4: Validation accuracy (%) on Scene-15 and MIT-67 with different learning rates and batch sizes.



(a) 1e-3 (b) 1e-4

Figure 6: Comparison of masks for two samples under learning rates of `1e-3` and `1e-4`.

Hyperparameter Analysis We assessed the model's sensitivity to hyperparameters by examining the effects of learning rate and batch size on the performance of GAMGN. Preliminary experiments identified a learning rate of `1e-5` as optimal for fine-tuning the ViT component. With the ViT learning rate fixed at `1e-5`, we first evaluated GAMGN across different learning rates (`1e-1`, `1e-2`, `1e-3`, and `1e-4`) while keeping the batch size fixed at 75. Subsequently, we analyzed the impact of batch size (75, 100, 125, and 150) while fixing the GAMGN learning rate at `1e-3`. All training was conducted for 300 epochs.

Table 4 reports validation accuracy under different learning rates and batch sizes on Scene-15 and MIT-67. As shown in Figure 6, increasing the learning rate sharpens the mask distribution and strengthens the model's focus on structural features, whereas excessively high values cause unstable optimization and degrade performance. Conversely, very low learning rates produce nearly uniform masks due to the CEL loss, reducing the model's ability to capture topological differences. To balance accuracy and mask quality, we set the GAT learning rate to `1e-3`, and use a batch size of 75, which offered stable and efficient training across datasets.

Config.	Acc. (%)	Params (M)	GMACs	Latency (ms)
ViT-B	84.18	85.8	17.57	3.88 ± 0.20
(16→8)	97.02	86.6	17.81	7.94 ± 0.32
(32→16)	98.58	88.7	18.05	8.17 ± 0.35
(64→32)	98.46	95.6	18.42	9.36 ± 0.41

Table 5: Computational complexity and accuracy of TANSR under different configurations under the standard inference protocol, compared with ViT-B.

Computational Complexity To further analyze the efficiency of TANSR, we examine how the architecture of GAMGN influences both performance and computational cost. Three configurations of GAT layers, denoted as `(a→b)` for two-layer settings with hidden dimensions a and b, were evaluated on MIT-67. The number of parameters and inference latency were measured under a standardized protocol (224×224 input, batch size 1, FP32 precision, PyTorch 2.5 + cuDNN, single RTX 4090).

As shown in Table 5, enlarging the model increases parameters, GMACs (billion multiply–accumulate operations), and latency while generally improving accuracy. Conversely, reducing the model size lowers computational cost but leads to a clear performance drop. Hence, the `(32→16)` configuration achieves the optimal trade-off between performance and complexity.

Overall, TANSR introduces moderate overhead relative to ViT-B, with an increase of approximately 3.4% in parameters, 2.7% in GMACs and over 12.8% in accuracy. The observed latency of 8.17 ± 0.35 ms remains practical for real-time scene recognition. These results confirm that TANSR effectively enhances structural representation while maintaining an acceptable computational footprint.

Conclusion

In this paper, we propose TANSR, a novel topology-aware attention network that integrates GGM, GAMGN and TAG to extract topological information from scenes and guide the attention mechanism of ViTs. Extensive experiments demonstrate that TANSR achieves SOTA performance on multiple widely used scene recognition benchmarks (Scene-15, MIT-67 and SUN397). Our findings highlight an effective multimodal fusion paradigm for transformers, where graph-derived features generate label-specific masks that weight patch embeddings, significantly enhancing transformer's capability in complex scene recognition tasks, particularly in visually similar yet structurally distinct scenes.

However, several challenges remain. The performance of the model inherently depends on the quality of the built scene graph, and the incorporation of additional graph processing introduces computational overhead, which limits real-time applications. Additionally, the use of topology-aware masks may also reduce the interpretability of the resulting attention maps. Future work will aim to address these limitations by exploring lightweight and efficient graph representations, improving graph generation quality, and enhancing the interpretability of topology-aware attention.

Acknowledgments

This work is supported in part by National Key Research and Development Program of China (No. 2024YFC2310801), National Natural Science Foundation of China (No. 62476052), Shenzhen Science and Technology Program (Nos. JCYJ20230807115959041 and JCYJ20230807120010021), and the Open Fund of the Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education (No. CBDIS202501).

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11): 2274–2282.

Chen, G.; Song, X.; Zeng, H.; and Jiang, S. 2020. Scene recognition with prototype-agnostic scene layout. *TIP*, 29: 5877–5888.

Cheng, X.; Lu, J.; Feng, J.; Yuan, B.; and Zhou, J. 2018. Scene recognition with objectness. *PR*, 74: 474–487.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, 886–893. Ieee.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hayat, M.; Khan, S. H.; Bennamoun, M.; and An, S. 2016. A spatial layout and scale invariant feature representation for indoor scene classification. *TIP*, 25(10): 4829–4841.

Herranz, L.; Jiang, S.; and Li, X. 2016. Scene recognition with cnns: objects, scales and dataset bias. In *CVPR*, 571–579.

Hsu, C.-Y.; and Ding, J.-J. 2013. Efficient image segmentation algorithm using SLIC superpixels and boundary-focused region merging. In *ICICSP*, 1–5. IEEE.

Irving, B. 2016. maskSLIC: regional superpixel generation with application to local pathology characterisation in medical images. *arXiv preprint arXiv:1606.09518*.

Lazebnik, S.; Schmid, C.; and Ponce, J. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, 2169–2178. IEEE.

Lin, C.; Lee, F.; Xie, L.; Cai, J.; Chen, H.; Liu, L.; and Chen, Q. 2022. Scene recognition using multiple representation network. *Appl. Soft Comput.*, 118: 108530.

Liu, Y.; Chen, Q.; Chen, W.; and Wassell, I. 2018. Dictionary learning inspired deep network for scene recognition. In *AAAI*, volume 32.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV*, 60: 91–110.

Lu, S.; Mahadevan, V.; and Vasconcelos, N. 2014. Learning optimal seeds for diffusion-based salient object detection. In *CVPR*, 2790–2797.

Niu, J.; Ma, X.; and Li, R. 2024. SC-ViT: Semantic Contrast Vision Transformer for Scene Recognition. In *IJCNN*, 1–8. IEEE.

Quattoni, A.; and Torralba, A. 2009. Recognizing indoor scenes. In *CVPR*, 413–420. IEEE.

Said, Y.; Atri, M.; Albahar, M. A.; Ben Atitallah, A.; and Alsariera, Y. A. 2023. Scene recognition for visually-impaired people's navigation assistance based on vision transformer with dual multiscale attention. *Mathematics*, 11(5): 1127.

Saleknia, A. H.; Bagheri, E.; Barshooi, A. H.; and Ayatollahi, A. 2024. NEM: Nested Ensemble Model for scene recognition. In *MVIP*, 1–6. IEEE.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 12026–12035.

Suzuki, S.; Ueda, N.; and Sklansky, J. 1993. Graph-based thinning for binary images. *Int. J. Pattern Recognit Artif Intell.*, 7(05): 1009–1030.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 6105–6114. PMLR.

Tarabalka, Y.; Benediktsson, J. A.; Chanussot, J.; and Tilton, J. C. 2010. Multiple spectral–spatial classification approach for hyperspectral data. *TGRS*, 48(11): 4122–4132.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*, 10347–10357. PMLR.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Wightman, R. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models. Accessed: December 18, 2024.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 3485–3492. IEEE.

Xie, G.-S.; Zhang, X.-Y.; Yan, S.; and Liu, C.-L. 2015. Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. *TCSVT*, 27(6): 1263–1274.

Xie, L.; Lee, F.; Liu, L.; Kotani, K.; and Chen, Q. 2020. Scene recognition: A comprehensive survey. *PR*, 102: 107205.

Yang, S.; and Ramanan, D. 2015. Multi-scale recognition with DAG-CNNs. In *ICCV*, 1215–1223.