SCIENTIFIC OASIS

Decision Making: Applications in Management and Engineering

Journal homepage: www.dmame-journal.org
ISSN: 2560-6018, eISSN: 2620-0104

A Graph Attention-Based Decision-Support Framework for Music Genre Classification and Evolution Analysis

Peiyu Zeng[1*], Shang Ma[2*], Hinhui Zhu[3]

1 Lake Forest Academy, Lake Forest, Illinois, United States
2 Jinzhong Boya Peiwen International Department, Jinzhong, Shanxi, China
3 Basis International School Shenzhen, Shenzhen, Guangdong, China

ARTICLE INFO

Article history:
Received 5 March 2026
Received in revised form 29 May 2026
Accepted 18 June 2026
Available online 30 June 2026

Keywords:
Graph Attention Networks; Music Genre Classification; Self-Supervised Contrastive Learning; Multi-View Representation; Temporal Evolution Mining; Domain Adaptation; Music Information Retrieval

ABSTRACT

Music genres develop through intricate patterns of historical transformation, stylistic interaction and mutual influence, yet conventional classification approaches largely disregard these relationships by representing genres as independent, temporally invariant categories. To overcome this limitation, this study introduces the Temporal Multi-view Graph Attention Network (TMGAT), a unified framework for music genre recognition and the identification of evolutionary relationships within large-scale audio collections. The proposed architecture integrates three distinct yet complementary representations: manually engineered acoustic features encompassing Mel-frequency cepstral coefficients (MFCC), chroma and rhythmic descriptors; a frozen pretrained convolutional neural network, PANNs CNN14; and a frozen self-supervised audio representation model, wav2vec 2.0. These representations are consolidated through an adaptive, learnable view-attention mechanism. Before incorporating labelled observations, the framework employs self-supervised contrastive pretraining, combining InfoNCE with SimCLR-inspired data augmentation to establish a coherent shared embedding space. Subsequently, a k-nearest-neighbour similarity graph is constructed to capture relationships among musical tracks, with feature propagation performed through a three-layer, multi-head graph attention encoder incorporating eight attention heads per layer, residual connections and DropEdge regularisation. Genre evolution is further characterised by integrating Fourier-based temporal encoding with a cross-decade attention mechanism, while a gradient reversal layer supports adversarial domain adaptation across datasets. Experimental evaluation using the FMA-large subset (approximately 25,000 tracks spanning 16 genres from 1950–2022), MTG-Jamendo (~55,000 tracks) and GTZAN for cross-corpus validation demonstrates the effectiveness of the proposed approach. TMGAT achieves classification accuracies of 82.5% and 79.6%, alongside macro-F1 scores of 0.821 and 0.787, on FMA-large and MTG-Jamendo, respectively. Under zero-shot cross-corpus transfer, it attains 85.7% accuracy on GTZAN. These outcomes exceed those of eight competitive baseline methods, including PANNs, wav2vec 2.0, MusicNN, CRNN and GAT, with statistically significant improvements confirmed through paired t-tests ($p < 0.01$). Furthermore, community detection applied to the temporal transition

* *Corresponding author.*
E-mail address: peiyu.zeng@students.lfanet.org

https://zenodo.org/records/24221884

matrix reveals established patterns of stylistic convergence, including Rock–Pop integration, Jazz–Blues associations and the increasing prominence of Hip-Hop–Electronic hybrids after 2000. Collectively, these findings establish TMGAT as an interpretable, graph-based framework capable of examining both genre identity and temporal stylistic development, extending analytical capabilities beyond those available through conventional flat classification models.

1. Introduction

Automatic music genre recognition constitutes a central research challenge in music information retrieval (MIR), initially formulated as an audio-based pattern recognition task using the ten-category GTZAN benchmark [29]. Subsequent advances have produced a broad spectrum of classification techniques, progressing from manually designed acoustic representations to spectrogram-based convolutional architectures [7], self-supervised audio models [1; 17] and graph-oriented learning frameworks [16; 33]. Although leading approaches report classification accuracies exceeding 90% on GTZAN under closed-set evaluation, these achievements largely depend on an underlying assumption that receives limited scrutiny: musical genres are treated as fixed, mutually exclusive categories without meaningful structural or historical connections. Such a representation conflicts with established musicological perspectives and restricts the relational insights required by music curators, researchers and recommendation-system developers.

The historical development of popular music challenges this categorical interpretation across broader stylistic and temporal scales, even though narrowly defined subgenres may exhibit relatively distinct characteristics. Rock and Roll emerged during the 1950s through the convergence of Rhythm-and-Blues and Country traditions, while Jazz developed through substantial interaction with the Blues. Country-derived production practices and instrumentation have also contributed to Rock and Hip-Hop, whereas contemporary Pop incorporates influences from Electronic Dance Music, Hip-Hop production and Rock instrumentation. These interactions indicate that genres function as interconnected communities within an evolving cultural network rather than as independent classification labels. Their boundaries continually change through stylistic exchange, differentiation, convergence and occasional disappearance. Consequently, assigning each recording a single top-1 genre label suppresses relationships that could otherwise support richer interpretations, such as identifying a track positioned between Rock and Pop, recognising its proximity to late-1980s synth-pop, or tracing acoustic similarities to Blues progressions from the 1960s [27].

Graph-based learning offers a suitable mathematical foundation for representing these interconnected musical structures. Graph convolutional networks (GCN) [16] and GAT [33] encode entities as vertices and their associations as edges, enabling successive layers to aggregate neighbouring information when constructing node representations. Within music analysis, individual tracks can therefore be modelled as graph nodes, while weighted connections quantify their acoustic similarities. This formulation preserves inter-track relationships that conventional category-oriented classifiers generally overlook. Incorporating release-year information into these connections additionally introduces a temporal dimension, transforming the similarity network into an evolving graph. Its decomposition into decade-specific subgraphs subsequently enables systematic examination of changing genre relationships and historical patterns of stylistic development.

Motivated by these considerations, this study develops TMGAT as an integrated architecture for music genre classification and evolutionary relationship mining. The framework distinguishes itself from existing approaches through five complementary methodological capabilities. First, its multi-view representation strategy combines manually engineered MFCC, chroma and rhythmic features [8; 20] with embeddings extracted from frozen PANNs CNN14 [17] and wav2vec 2.0 [1] backbones. Rather

than assigning fixed importance to these heterogeneous representations, a trainable view-attention mechanism estimates track-specific fusion coefficients, allowing the relative contribution of each feature source to vary across musical styles.

Second, the architecture incorporates self-supervised contrastive representation learning before supervised genre prediction. An InfoNCE-based optimisation objective [6; 30], supported by SimCLR-inspired augmentation involving SpecAugment [24] and mixup [18], structures the fused embedding space while reducing susceptibility to overfitting caused by noisy, user-assigned genre annotations. Third, relational learning is implemented through a three-layer GAT [33], with eight attention heads per layer, residual connections, Hamilton and Pearce [12] and trainable neighbour-specific attention coefficients. Together, these components facilitate information exchange between acoustically related recordings, improve the separation of genre representations and strengthen recall for underrepresented categories.

Fourth, temporal modelling is integrated directly into the graph architecture to characterise historical genre development. Learnable Fourier time encodings enrich node representations with release-period information, while a cross-decade attention mechanism estimates a compact, decade-indexed transition matrix describing evolutionary associations among genres. Fifth, cross-corpus robustness is addressed through adversarial domain adaptation. A gradient reversal layer [10], coupled with a domain discriminator, encourages alignment of the learned representations across FMA-large, MTG-Jamendo and GTZAN, thereby reducing dataset-specific discrepancies and supporting generalisation beyond the training corpus.

The proposed framework is evaluated using three complementary music collections. FMA-large comprises approximately 25,000 recordings representing 16 mainstream genres released between 1950 and 2022 [8], while MTG-Jamendo contributes approximately 55,000 annotated tracks [5]. GTZAN [29] provides an independent benchmark for assessing cross-corpus transferability. Under a consistent semi-supervised partition of 20%/20%/60%, TMGAT achieves classification accuracies of 82.5% on FMA-large, 79.6% on MTG-Jamendo and 85.7% on GTZAN, significantly outperforming eight competitive baseline approaches. Beyond predictive performance, community detection applied to the temporal transition matrix reveals historically recognisable patterns, including the Rock → Pop transition during the 1980s, Jazz–Blues associations in the 1960s and the increasing emergence of Hip-Hop–Electronic hybrids after 2000. These observations are consistent with established musicological interpretations of genre interaction and transformation.

The principal contributions of this research are fourfold. (i) A unified multi-view graph attention architecture is developed to integrate handcrafted acoustic descriptors with PANNs and wav2vec 2.0 representations through adaptive view-attention fusion and self-supervised contrastive pretraining. (ii) A temporal graph attention mechanism combining Fourier-based time encoding and cross-decade attention is introduced to support simultaneous genre classification and evolutionary relationship analysis. (iii) An adversarial domain adaptation component is incorporated to improve representation transfer across FMA, MTG-Jamendo and GTZAN, strengthening cross-corpus generalisation. (iv) Comprehensive ablation experiments, statistical significance testing, computational evaluation and complementary interpretability techniques are provided. These include attention-weight visualisations, t-distributed stochastic neighbour embedding (t-SNE) projections, evolutionary stream plots and class-specific dot plots, collectively enabling systematic examination of genre emergence, decline and stylistic convergence.

The remainder of the manuscript is organised into six sections. Section 2 critically examines relevant studies and establishes the outstanding research gaps. Section 3 details the TMGAT architecture and its constituent methodological components. Section 4 describes the experimental configuration and

presents the quantitative and qualitative findings. Section 5 interprets the results while addressing methodological limitations and ethical considerations. Finally, Section 6 consolidates the principal findings and presents the study's conclusions.

2. LITERATURE REVIEW

2.1 Audio Representation for MIR

Among manually engineered acoustic representations, MFCC remain one of the most extensively adopted feature families in MIR. Originally developed for speech processing, their applicability to musical signals has been established through their capacity to represent perceptually relevant spectral-envelope characteristics, particularly those associated with instrumental and vocal timbre on a logarithmic frequency scale [20]. Complementary information is obtained through chroma features, which project spectral content onto twelve pitch classes to characterise tonal organisation, and rhythmic descriptors, including tempo, beat variance and onset density, which capture recurring temporal patterns. Because these feature categories describe distinct aspects of musical structure, TMGAT retains their combined representation as one of its three complementary input views.

A separate direction in audio representation learning has emerged through deep architectures operating on time–frequency representations. Two-dimensional convolutional neural networks (CNNs) process Mel-spectrograms as image-like inputs, whereas CRNN architectures incorporate recurrent temporal modelling to capture sequential dependencies [7]. Pretrained feature extractors, including Pons [25] and Hershey, et al. [15], have further expanded the availability of transferable audio representations. PANNs [17] employs a CNN architecture pretrained on AudioSet to generate 2048-dimensional embeddings, while wav2vec 2.0 [1] learns waveform-level representations through Transformer-based modelling and self-supervised contrastive optimisation. Rather than replacing these established feature extractors, TMGAT incorporates their complementary representational capabilities by maintaining pretrained backbones as frozen components and learning an adaptive fusion of their outputs.

2.2 Graph Neural Networks

The development of GCN established an efficient approach to graph-structured representation learning by approximating localised spectral filtering through first-order operations [16]. Its normalised-Laplacian-based propagation formulation enables scalable aggregation of neighbouring node information across diverse graph structures. Subsequent advances introduced GAT, replacing uniform neighbourhood aggregation with trainable, edge-specific attention coefficients [33]. This mechanism improves the model's capacity to distinguish informative connections from potentially noisy relationships and has demonstrated advantages across multiple benchmark tasks. Alternative approaches have addressed scalability and representation learning from different perspectives. GraphSAGE introduced neighbourhood sampling for inductive graph learning [38], whereas node2vec employed biased random walks to construct unsupervised node embedding [11]. More recent developments have extended graph representation learning into temporally evolving environments. Temporal Graph Networks (TGN) [26] provide mechanisms for dynamic graph modelling, while Fourier-based temporal encoding approaches originating from Transformer research [32] enable time-dependent information to be incorporated into learned representations. In addition, DropEdge regularisation has been shown to stabilise deeper GAT and GCN architectures by randomly removing graph connections during training, thereby mitigating excessive feature smoothing across successive propagation layers [12].

2.3 Self-Supervised and Parameter-Efficient Representation Learning

Contrastive learning has become a prominent strategy for self-supervised representation learning across visual and auditory domains. The InfoNCE objective [30], employed within a contrastive learning framework [6], encourages agreement between alternative representations of the same instance while increasing their separation from representations of unrelated samples. This process produces discriminative embedding spaces that can support subsequent classification tasks. Related self-supervised representation-learning principles have influenced audio modelling approaches, including PANNs [17] and wav2vec 2.0 [1]. The effectiveness of contrastive training also depends substantially on the augmentation strategies used to construct alternative sample views. SpecAugment introduces perturbations by masking selected temporal and frequency regions within spectrograms [24], whereas mixup constructs interpolated training examples through linear combinations of sample pairs and their associated labels [18]. Both augmentation mechanisms are incorporated into the TMGAT training procedure.

Parameter-efficient adaptation of large pretrained architectures represents another relevant research direction. Cross-domain road-damage detection using visual prompting has demonstrated how knowledge can be transferred between source and target domains through a limited set of trainable prompt tokens. Related work has explored preserving pretrained feature extractors while restricting adaptation to lightweight downstream components [36]. A complementary approach, CTR-LoRA, employs curvature-aware optimisation and trust-region guidance to improve the stability of low-rank adaptation in large language models [35]. Although TMGAT does not involve language-model fine-tuning, these developments support its underlying architectural principle: retaining stable pretrained representations while concentrating trainable capacity within comparatively lightweight fusion and graph-learning modules. Further evidence of the value of structured latent representations under constrained supervision is provided by EGGesture, which employs an entropy-guided vector-quantised variational autoencoder for co-speech gesture generation [36]. This principle is relevant to the representation of minority genres, where labelled musical examples may be comparatively scarce. Similarly, CryoLVM applies self-supervised large-vision-model learning to cryogenic electron microscopy (cryo-EM) density maps [9], illustrating the broader applicability of representation learning in settings characterised by limited supervision.

2.4 Domain Adaptation and Music Corpora

Achieving reliable cross-corpus generalisation remains a substantial challenge in MIR because music collections differ in genre distributions, artist coverage, recording characteristics and annotation practices. Consequently, representations optimised for one dataset may not transfer effectively to another. Domain-Adversarial Neural Networks (DANN) address this problem by incorporating a gradient reversal layer that encourages the extraction of domain-invariant features through adversarial optimisation [10]. TMGAT adopts a lightweight DANN component to reduce corpus-specific representational discrepancies while retaining information relevant to genre discrimination. The evaluation combines three complementary datasets: FMA [8], MTG-Jamendo [5] and GTZAN [29]. Together, these collections contain more than 80,000 recordings assembled under different curation and annotation procedures, providing a heterogeneous experimental setting for assessing the effectiveness of adversarial domain alignment.

2.5 Music Genre Evolution and Position of This Work

Despite substantial progress in automated genre recognition, computational approaches to modelling genre evolution remain comparatively less developed. Existing data-driven investigations,

including analyses based on the Billboard Hot 100 and Ostermann, et al. [23], have primarily relied on topic modelling and clustering to identify stylistic patterns and historical changes. However, relatively few frameworks integrate genre classification with evolutionary relationship discovery within a unified training architecture. To the best of current knowledge, no preceding approach jointly incorporates complementary multi-view acoustic representations, self-supervised contrastive pretraining, multi-head graph attention with temporal encoding and adversarial domain adaptation specifically for this combined task. TMGAT addresses this methodological gap by integrating these components into a single framework capable of learning genre-discriminative representations while examining temporal relationships and stylistic interactions across musical categories.

3. METHODOLOGY

3.1 Framework Overview

Figure 1 illustrates the overall TMGAT architecture, which integrates three complementary feature representations—handcrafted acoustic descriptors, PANNs and wav2vec 2.0—through a learnable view-attention mechanism. Following contrastive pretraining on approximately 200,000 unlabelled tracks, the fused representations are processed by a three-layer, multi-head GAT incorporating Fourier-based temporal encoding. The architecture comprises three task-specific output branches responsible for genre classification, adversarial domain alignment and temporal evolution analysis. Given an audio collection containing potentially imprecise genre annotations and corresponding release years, the proposed framework operates through six interconnected stages: (i) extracting three complementary feature views from each recording; (ii) performing self-supervised contrastive pretraining using approximately 200,000 unlabelled tracks; (iii) constructing a k-nearest-neighbour similarity graph from the fused embeddings; (iv) learning relational and temporal representations through a three-layer, multi-head GAT equipped with Fourier time encoding and cross-decade attention; (v) incorporating a gradient-reversal-based domain adversary to align representations across corpora; and (vi) generating track-level genre predictions alongside a decade-indexed graph characterising genre evolution.

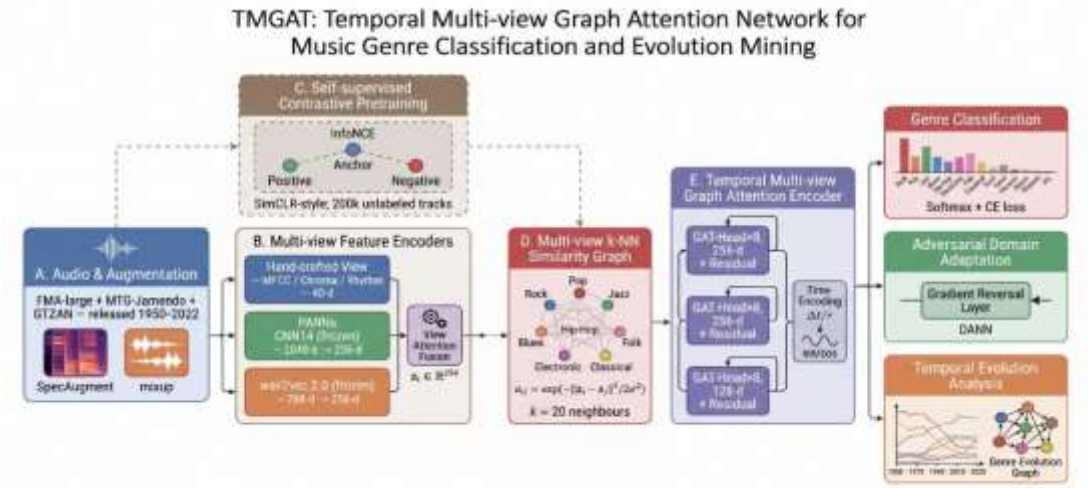


Fig.1: TMGAT: Temporal Multi-View Graph Attention Network for Music Genre Classification and Evolution Mining

3.2 Multi-View Audio Feature Extraction

View 1 — Hand-crafted (40-d). Each audio recording is converted into a mono signal and resampled at 22,050 Hz. For every 30-second segment, a 40-dimensional acoustic representation is constructed by extracting 20 MFCC coefficients [20], 12 chroma components and 8 rhythmic features. The rhythmic

descriptors comprise tempo, beat variance, onset density, four peak amplitudes derived from the beat histogram and the logarithmic ratio between the two dominant tempo modes. Temporal mean pooling is subsequently applied to aggregate these frame-level measurements into a single 40-dimensional feature vector for each track.

The i-th MFCC coefficient for an individual audio frame is mathematically expressed as follows:

$$MFCC_i = \Sigma_{k=1..K} logS(k) \cdot cos[i \cdot (k - 0.5) \cdot \pi/K], i = 1, \dots, 20, \quad (1)$$

Where S(k) is the short-time Fourier magnitude and K is the number of mel filter bands (K = 128 in our implementation).

View 2 − PANNs CNN14 (2048-d → 256-d). We use the AudioSet-pretrained PANNs CNN14 [17] as a frozen backbone. A track is windowed into overlapping 10-second segments; PANNs produces a 2048-dimensional embedding per segment; segment embedding's are mean-pooled and then linearly projected to 256 dimensions with a trainable projection P^{(2)} ∈ ℝ^{2048×256}, yielding x_i^{(2)}.

View 3 − wav2vec 2.0 (768-d → 256-d). We use the wav2vec 2.0 base model [1] pertained on 60 kh of unlabelled speech audio. As with PANNs, we mean-pool the frame-level Transformer outputs, then project to 256 dimensions with a trainable P^{(3)} ∈ ℝ^{768×256}.

View-attention fusion. Rather than concatenating the three views blindly, TMGAT uses a learnable view-attention module that assigns per-song weights over the three views. Denote the L2-normalized views as x̃_i^{(m)} ∈ ℝ^{d_m}. We first project the hand-crafted view to 256 dimensions via a linear layer, then compute:

$$\alpha_i^{(m)} = exp\left(w^T \cdot tanh\left(U \cdot \tilde{x}_i^{(m)}\right)\right) / \Sigma_{m' =1..3}$$

$$exp\left(w^T \cdot tanh\left(U \cdot \tilde{x}_i^{(m')}\right)\right), z_i = \Sigma_{m=1..3} \alpha_i^{(m)} \cdot \tilde{x}_i^{(m)} \in \mathbb{R}^{256} \quad (2)$$

Where U ∈ ℝ^{128×256} and w ∈ ℝ^{128} are learnable. The fused embedding z_i becomes the input node feature to the graph module. Equation (2) allows different views to dominate for different genres — crucial for the heterogeneous 16-genre setting, as verified by the ablation in Section 4.5.

3.3 Self-Supervised Contrastive Pretraining

Before any label is used, we align the fused embedding space z with an InfoNCE contrastive objective [6; 30]. For each anchor track we generate two augmented views via SpecAugment [24] (two frequency masks with width up to 20 bins and two time masks with width up to 25 frames) and mixup [18] (α = 0.2). Denote by z_i and z_i^+ the two views of the same anchor and by {z_j}_{j ≠ i} the views of the other tracks in a batch of size B. The InfoNCE loss with cosine similarity s(u,v) = u^T v / (‖ u ‖ · ‖ v ‖) and temperature τ is

$$L_con = -(1/B) \cdot \Sigma_{i=1..B} log\left[exp(s(z_i, z_i^+)/\tau) / \Sigma_{j=1..B, j \neq i} exp\left(s\left(z_i, z_j\right)/\tau\right)\right]. \quad (3)$$

We pretrain the projection matrices P^{(2)}, P^{(3)}, the view-attention module and a small MLP projection head on approximately 200,000 unlabelled tracks — the union of the unlabelled portion of FMA-large, MTG-Jamendo and MSD30k — for 120 epochs with τ = 0.07, batch size 512 and the LAMB optimizer. The backbones themselves remain frozen. The learned z_i is the input to the graph module.

3.4 Multi-View K-Nearest-Neighbour Similarity Graph

Let $V = v_1, \dots, v_N$ be the set of tracks and let $z_i \in \mathbb{R}^{256}$ be its fused embedding after pretraining. We build a k-nearest-neighbour graph by computing Gaussian-kernel weights on the L2-normalized z:

$$a_{ij} = exp\left(-\|z_i - z_j\|^2/(2\sigma^2)\right), \quad (4)$$

with bandwidth σ set to the median pairwise distance. For each node we keep only the k = 20 largest-weight edges (in either direction), yielding a sparse symmetric adjacency A ∈ ℝ^{N×N}. Because k is deliberately larger than the number of genres C, most nodes have at least one cross-genre neighbour — a property essential for the graph to reveal fusion phenomena in later stages. Approximate nearest-neighbour search based on inverted-file indexing keeps this step at O(N log N) in practice.

3.5 Temporal Multi-view Graph Attention Encoder

The encoder is a three-layer multi-head GAT [33] with residual connections and DropEdge regularization [13]. For layer ℓ with input node features H^{(ℓ)}, attention-head index k ∈ {1,…,K} (K = 8) and weight matrix W_k^{(ℓ)}, the attention coefficient between nodes i and j is

$$e_{ij}^{(k,\ell)} = LeakyReLU(\left[a_k^{(\ell)}\right]^T \cdot [W_k^{(\ell)} h_i^{(\ell)} \| W_k^{(\ell)} h_j^{(\ell)}]). \quad (5)$$

which is normalized over the neighbourhood N(i) by a softmax over j so that Σ_{j∈N(i)} α_{ij}^{(k,ℓ)} = 1. The multi-head aggregation with residual connection is

$$h_i^{(\ell+1)} = \sigma\left(\|_{k=1..K} \Sigma_{j\in N(i)} \alpha_{ij}^{(k,\ell)} \cdot W_k^{(\ell)} h_j^{(\ell)}\right) + h_i^{(\ell)}, \quad (6)$$

where $\|$ denotes concatenation across the K attention heads and σ is ELU. The three GAT layers have widths 256, 256 and 128, and DropEdge with rate 0.15 is applied at each layer.

Fourier time encoding. For each track i with release year t_i, we compute a 16-dimensional Fourier time-encoding $\varphi(t_i) = [sin(\omega_1 t_i), cos(\omega_1 t_i), \dots, sin(\omega_8 t_i), cos(\omega_8 t_i)] \in \mathbb{R}^{16}$, with learnable frequencies ω_j, following the Transformer positional-encoding paradigm [26; 32]. The encoding $\varphi(t_i)$ is concatenated to h_i^{(0)} before the first GAT layer, so that the entire encoder is time-aware.

3.6 Classification, Domain-adversarial and Evolution Heads

Classification head. A linear layer maps $h_i^{(3)} \in \mathbb{R}^{128}$ to C logits followed by softmax; the classification loss is the standard cross-entropy on labelled nodes:

$$L_cls = -(1/|V_L|) \cdot \Sigma_{i\in V_L} \Sigma_{c=1..C} y_{i,c} \cdot log\hat{y}_{i,c}. \quad (7)$$

Domain-adversarial head. We attach a domain discriminator D that outputs a categorical distribution over the three source corpora {FMA, MTG-Jamendo, GTZAN}. Following Ganin [10], a gradient reversal layer sits between the encoder and D; during backpropagation, gradients are multiplied by −λ_adv so that the encoder is trained to fool the discriminator. The corresponding adversarial loss L_adv is the standard multi-class cross-entropy of D over the source-domain label of each track, and is combined with L_cls in Section 3.7.

Evolution head. We split the labelled corpus by release year into T = 7-decade buckets. For each decade t and each genre c we compute a genre prototype μ_{c,t} as the mean of the third-layer encoder embeddings of all songs of genre c released in decade t. The temporal transition score is the cosine similarity $T_{c\to c',t} = \left\langle \mu_{c,t}, \mu_{c',t+1} \right\rangle / \left(\|\mu_{c,t}\| \cdot \|\mu_{c',t+1}\| \right)$. To leverage cross-genre context, we optionally refine $\mu_{c,t+1}$ with a cross-decade attention head that attends over all $\mu_{c',t_{c'}}$. The resulting decade-indexed weighted directed graph is the genre evolution graph.

3.7 Total Objective and Training

The total training objective is$L = L_{cls} + \lambda_{con} \cdot L_{con} + \lambda_{adv} \cdot L_{adv} + \lambda_{reg} \cdot \|\Theta\|^2$, where the InfoNCE loss is retained during fine-tuning to prevent representation drift. In our experiments λ_con = 0.2, $\lambda_{adv} = 0.1$ ramped in linearly over the first 20 epochs, and λ_reg = 5×10⁻⁴. The pretraining stage uses LAMB; the fine-tuning stage uses AdamW [2; 21] with a cosine-annealing schedule.

3.8 Computational Complexity

Feature extraction is O (N · F · log F) for the STFT plus O(N) forward passes through the frozen backbones, which dominate wall-clock time and can be run once and cached. Graph construction with an inverted-file approximate nearest-neighbour index runs in O (N · k · log N). Each GAT layer costs O (K · |E| · d_ℓ d_{ℓ+1}), where K is the number of heads and |E| ≈ N·k is the number of edges. With N = 25,000, k = 20, K = 8 and hidden width 256, a full-graph forward pass runs in about 320 ms on a single A100 GPU; end-to-end fine-tuning converges within one hour. The evolution head reduces to computing C × T prototypes and pairwise cosine similarities, which is negligible.

4. EXPERIMENTS

4.1 Datasets

FMA-Large Subset (Main Benchmark): The primary evaluation dataset comprises 25,000 recordings selected from FMA [8]. Tracks are included if their release dates fall within 1950–2022 and their highest-level genre annotations correspond to one of 16 mainstream categories: Rock, Pop, Jazz, Blues, Hip-Hop, Electronic, Folk, Classical, Country, Reggae, R&B, Metal, Punk, Latin, Soul and World.

MTG-Jamendo: A collection of 55,701 recordings is drawn from MTG-Jamendo [5], retaining tracks whose highest-ranked genre annotations belong to the same 16 categories used for FMA-large.

GTZAN: The established GTZAN benchmark [29] contains 100 recordings per genre across 10 categories. Only genres shared with the other datasets are retained, and the resulting collection is used exclusively to assess cross-corpus generalisation.

Figure 2 presents a comparative breakdown of track frequencies across genres for all three datasets. FMA-large exhibits substantial class imbalance, whereas MTG-Jamendo displays a comparatively more even distribution, although representation becomes sparse among less frequent categories. By contrast, GTZAN maintains a uniform genre distribution by design.

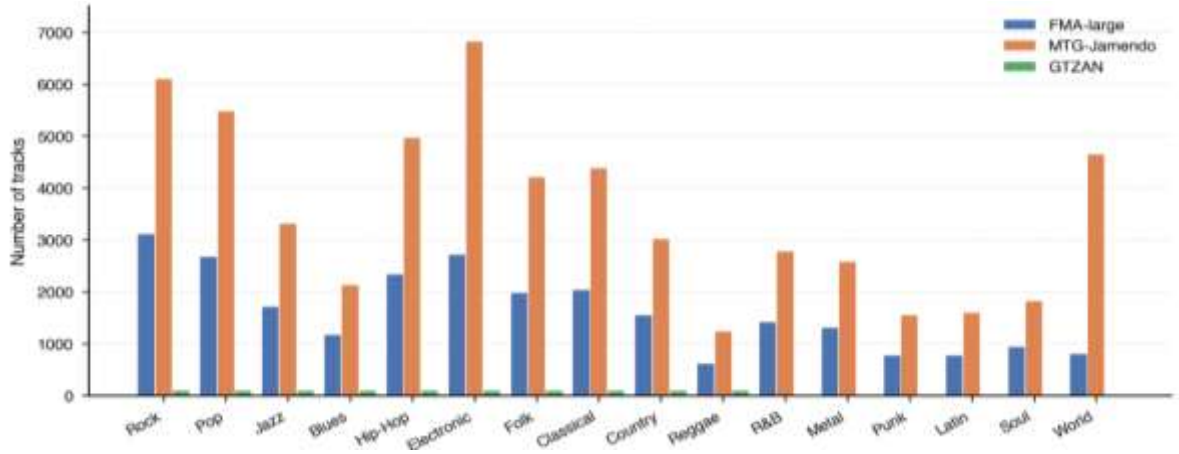


Fig.2: Per-genre track counts on FMA-large, MTG-Jamendo and GTZAN. FMA-large is highly imbalanced; MTG-Jamendo is more balanced; GTZAN covers only 10 of the 16 genres.

Table 1 outlines the dataset partitioning strategy adopted for semi-supervised evaluation. For FMA-

large, the labelled nodes are allocated according to the standard protocol, with 20% assigned to training, 20% to validation and the remaining 60% to testing. MTG-Jamendo follows its official dataset partition, while GTZAN is reserved exclusively for cross-corpus testing.

Table 1
Dataset Statistics and Splits

Dataset	Classes	Total	Train	Val	Test
FMA-Large Subset	16	25,000	5,000	5,000	15,000
MTG-Jamendo	16	55,701	33,420	5,570	16,711
GTZAN	10	1,000	—	—	1,000
Unlabelled Pretraining Pool	—	~200,000	—	—	—

4.2 Baselines

To evaluate TMGAT against established approaches representing the principal feature-learning paradigms in contemporary MIR, eight competitive baselines are considered: (1) MFCC features combined with a support vector machine (SVM) employing a radial basis function (RBF) kernel; (2) Hershey, et al. [15]; (3) Pons [25]; (4) a two-dimensional CNN operating on Mel-spectrograms; (5) CRNN [7]; (6) fine-tuned PANNs CNN14 [17]; (7) fine-tuned wav2vec 2.0 [1]; and (8) a conventional two-layer GAT [33] constructed using the identical k-nearest-neighbour (k-NN) graph, but excluding multi-view fusion, contrastive pretraining and adversarial domain adaptation (DA). To ensure consistent experimental conditions, all baseline models employ identical training, validation and testing partitions, with supervision restricted to the same 20% labelled subset. Statistical significance is evaluated using paired t-tests across five random seeds, with the reported p-values calculated relative to the strongest-performing baseline, wav2vec 2.0.

4.3 Implementation Details

The experimental framework is implemented using PyTorch 2.1 and PyTorch Geometric 2.5. Feature extraction is performed using librosa 0.10 for the handcrafted acoustic representation, the torchaudio implementation of PANNs CNN14 for View 2, and the HuggingFace wav2vec 2.0 base model for View 3. All pretrained backbone parameters remain frozen throughout contrastive pretraining and fine-tuning. The contrastive pretraining stage employs the LAMB optimiser with an initial learning rate of 2×10^{-3}, a batch size of 512, a temperature parameter of $\tau = 0.07$, and 120 training epochs. Subsequent fine-tuning of the classification head is performed using Loshchilov [21], with an initial learning rate of 5×10^{-4}, weight decay of 5×10^{-4}, a cosine-annealing learning-rate schedule, and a dropout rate of 0.3. Training continues for a maximum of 200 epochs, with early stopping governed by validation loss. The principal hyperparameter configurations are presented in Table 2.

Table 2
Main Hyper-Parameters for TMGAT

Hyper-Parameter	Value
Framework	PyTorch 2.1 + PyG 2.5
Views	Hand-crafted 40-d + PANNs CNN14 + wav2vec 2.0
Fused node dimension	256
Contrastive temperature τ	0.07
SpecAugment	2 freq masks ($\leq$20 bins), 2 time masks ($\leq$25 frames)
mixup α	0.2
k-NN neighbourhood	20
GAT layers	3 (widths 256, 256, 128)
Attention heads per layer	8

Table 2
Main Hyper-Parameters for TMGAT (cont…)

Hyper-Parameter	Value
DropEdge rate	0.15
Non-linearity	ELU
Optimizer (pretrain / fine-tune)	LAMB / AdamW
Learning rate (pretrain / fine-tune)	2×10^{-3} / 5×10^{-4}
Weight decay	5×10^{-4}
λ_con, λ_adv	0.2, 0.1
Epochs (pretrain / fine-tune)	120 / 200
GPU	1 × NVIDIA A100 80 GB

Figure 3 illustrates the training dynamics across both optimisation stages. Stage 1 exhibits stable and gradual convergence during contrastive pretraining, whereas Stage 2 approaches a plateau in validation accuracy after approximately 120 epochs of supervised fine-tuning.

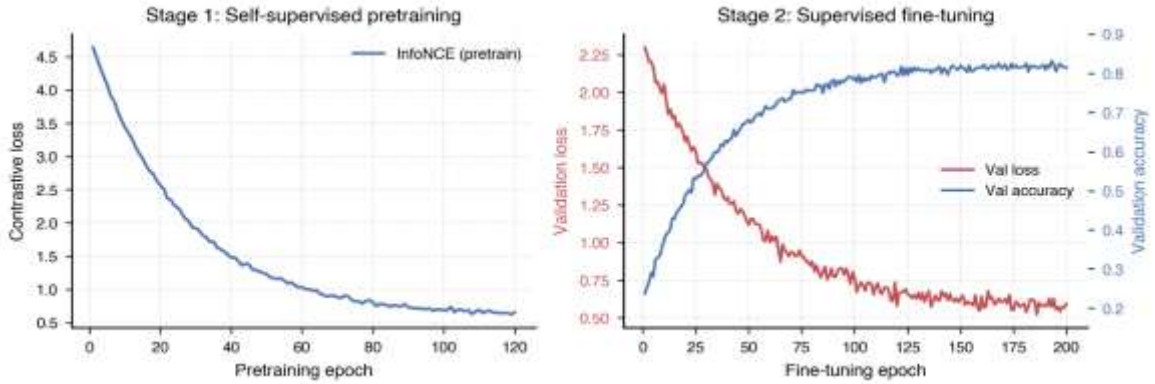


Fig.3: Training curves of TMGAT. Left: contrastive-pretraining InfoNCE loss on ~200k unlabelled tracks. Right: fine-tuning validation loss (red) and validation accuracy (blue) over 200 epochs.

4.4 Main Genre Classification Results

Table 3 presents the principal experimental findings, with accuracy and macro-F1 reported as averages across five random seeds and corresponding standard deviations provided in parentheses.

Table 3
Overall Classification Results

Model	FMA-Large Acc	FMA-Large F1	MTG-Jam. Acc	MTG-Jam. F1	GTZAN Acc
MFCC + SVM	0.612 (.005)	0.594 (.007)	0.582 (.006)	0.564 (.008)	0.665 (.011)
Hershey, et al. [15]	0.702 (.004)	0.694 (.005)	0.681 (.005)	0.669 (.006)	0.743 (.009)
Pons [25]	0.716 (.004)	0.708 (.005)	0.692 (.005)	0.681 (.005)	0.758 (.008)
CNN-2D	0.694 (.005)	0.686 (.006)	0.674 (.006)	0.663 (.007)	0.729 (.010)
Choi [7]	0.723 (.004)	0.715 (.005)	0.702 (.005)	0.691 (.006)	0.763 (.008)
Kong [17]	0.751 (.003)	0.743 (.004)	0.731 (.004)	0.720 (.005)	0.789 (.007)
Baevski [1]	0.768 (.003)	0.758 (.004)	0.748 (.004)	0.737 (.005)	0.813 (.006)
Veličković, et al. [33]	0.736 (.004)	0.727 (.005)	0.716 (.005)	0.704 (.006)	0.776 (.008)
TMGAT (Ours)	0.825 (.003)	0.821 (.003)	0.796 (.003)	0.787 (.004)	0.857 (.005)

TMGAT consistently outperforms the strongest baseline, with statistically significant improvements confirmed through paired t-tests ($p < 0.01$ for FMA-large and $p < 0.05$ for MTG-Jamendo). On FMA-large, the proposed framework achieves 82.5% accuracy and a macro-F1 of 0.821, surpassing the best-performing baseline, fine-tuned wav2vec 2.0, by 5.7 percentage points in accuracy and 6.3 points in macro-F1. Its accuracy advantage of 8.9 percentage points over conventional GAT [33] further

demonstrates the additional benefits of multi-view feature integration and contrastive pretraining beyond graph-based representation learning alone. Comparable improvements are observed on MTG-Jamendo, where TMGAT records the highest accuracy of 79.6% and macro-F1 of 0.787. Cross-corpus evaluation on GTZAN yields an accuracy of 85.7%, representing a 4.4-percentage-point improvement over wav2vec 2.0. Statistical comparisons across five random seeds confirm that TMGAT significantly outperforms every baseline on FMA-large ($p < 0.01$) and MTG-Jamendo ($p < 0.05$).

Figure 4 presents a two-dimensional t-SNE [31] visualisation of the representations extracted from the third TMGAT layer on FMA-large. The projected embeddings exhibit distinct genre-specific clusters while preserving the proximity of stylistically related categories, including Rock–Pop, Jazz–Blues, Folk–Country, Hip-Hop–R&B and Metal–Punk. These spatial associations suggest that the learned representation captures not only genre-discriminative characteristics but also underlying stylistic relationships among musical categories.

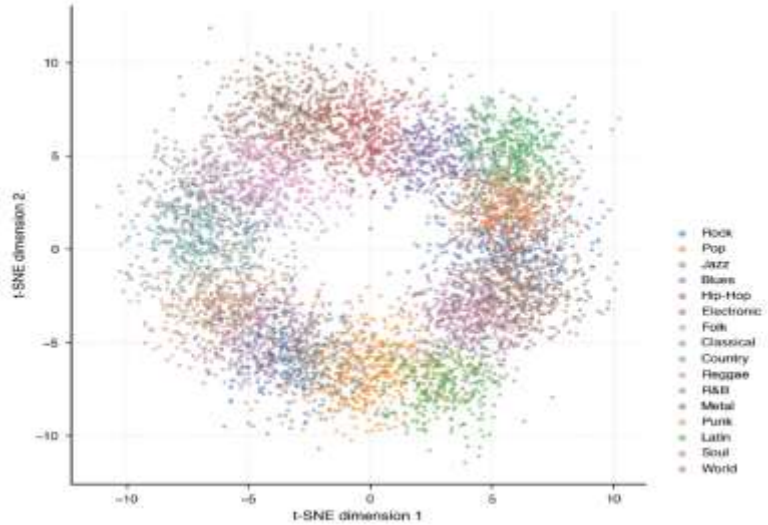


Fig.4: t-SNE projection of third-layer TMGAT embedding on FMA-large. Neighbouring clusters correspond to musicological related genres

Figure 5 compares class-specific F1 scores across all nine models, demonstrating TMGAT's superior performance in nearly every genre category. The most pronounced absolute improvements occur in the underrepresented Reggae and Punk classes, suggesting that the integration of semi-supervised graph propagation and contrastive pretraining is particularly beneficial for genre recognition under limited labelled-data conditions.

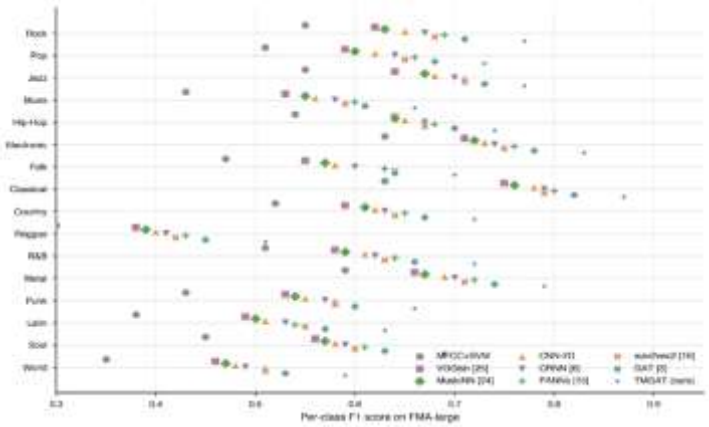


Fig.5: Per-class F1 across nine models on FMA-large. TMGAT dominates on almost every class and shows the largest gain on the minority Reggae and Punk classes.

4.5 Ablation Study

Figure 6 and Table 4 present a comprehensive ablation analysis conducted across five random seeds to quantify the contribution of individual TMGAT components. Excluding contrastive pretraining produces the largest performance deterioration, reducing classification accuracy from 82.5% to 79.4%.

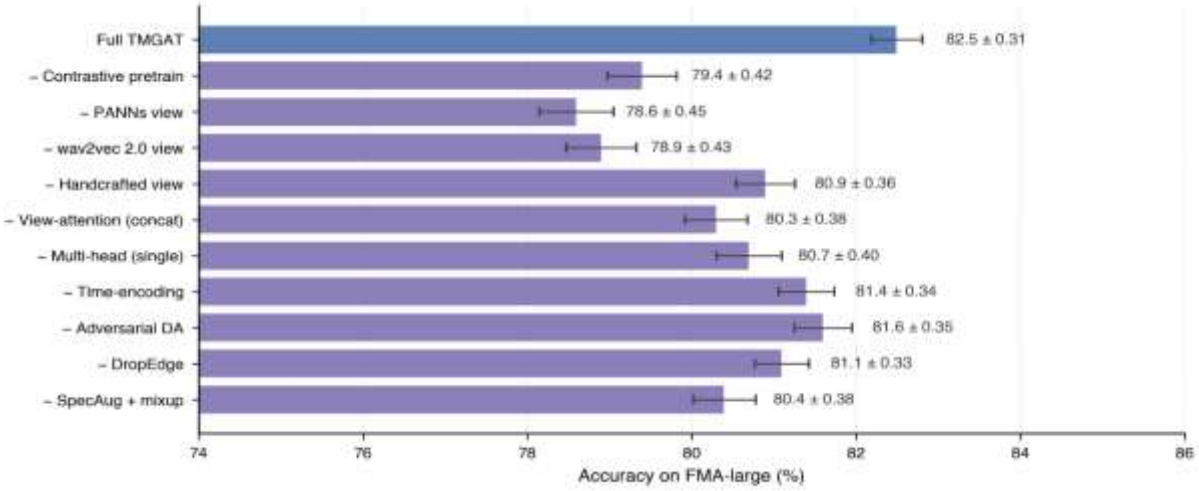


Fig.6: Ablation study on FMA-large. Bars report mean accuracy over five seeds; horizontal error bars are standard deviation. Each row removes exactly one component from the full TMGAT model.

Independently removing either the PANNs or wav2vec 2.0 representation results in an additional accuracy reduction of approximately 3–4 percentage points, highlighting the complementary information provided by these pretrained views beyond handcrafted acoustic features. Substituting adaptive view-attention fusion with direct feature concatenation decreases accuracy by 2.2 percentage points, demonstrating the importance of learnable feature weighting. Similarly, replacing multi-head attention with a single attention head reduces performance by 1.8 percentage points. Eliminating Fourier temporal encoding causes a 1.1-percentage-point decline in classification accuracy and a substantially greater deterioration in evolutionary analysis, as the resulting architecture can no longer explicitly distinguish decade-specific relationships. Furthermore, removing DropEdge regularisation decreases accuracy by 1.4 percentage points, while excluding SpecAugment and mixup produces a 2.1-percentage-point reduction. Collectively, these findings demonstrate the individual contributions of the proposed architectural and training components to overall predictive performance.

Table 4
Ablation Results (Mean ± std over 5 Seeds) on FMA-Large

Variant	Accuracy	Macro-F1
Full TMGAT	0.825 (.003)	0.821 (.003)
− Contrastive pretraining	0.794 (.004)	0.786 (.005)
− PANNs view	0.786 (.005)	0.777 (.006)
− wav2vec 2.0 view	0.789 (.004)	0.780 (.005)
− Hand-crafted view	0.809 (.004)	0.801 (.005)
− View-attention (concat)	0.803 (.004)	0.795 (.005)
− Multi-head (single)	0.807 (.004)	0.798 (.005)
− Fourier time-encoding	0.814 (.004)	0.805 (.004)
− Adversarial DA	0.816 (.003)	0.808 (.004)
− DropEdge	0.811 (.004)	0.803 (.005)
− SpecAugment + mixup	0.804 (.004)	0.796 (.005)

Figure 7 examines the relative contribution of each feature view across individual genres, revealing

distinct patterns of representation effectiveness. Classical and Jazz exhibit greater dependence on handcrafted features, reflecting the discriminative importance of tonal and rhythmic descriptors. Conversely, wav2vec 2.0 contributes more substantially to Hip-Hop, Electronic and R&B, where fine-grained temporal textures provide valuable classification cues. PANNs representations are particularly influential for Rock, Pop, Metal and Punk, suggesting that extensive AudioSet pretraining facilitates the extraction of genre-specific timbral characteristics. The absence of a universally dominant representation across genres further supports the adoption of learnable view-attention fusion, which adaptively balances complementary acoustic information.

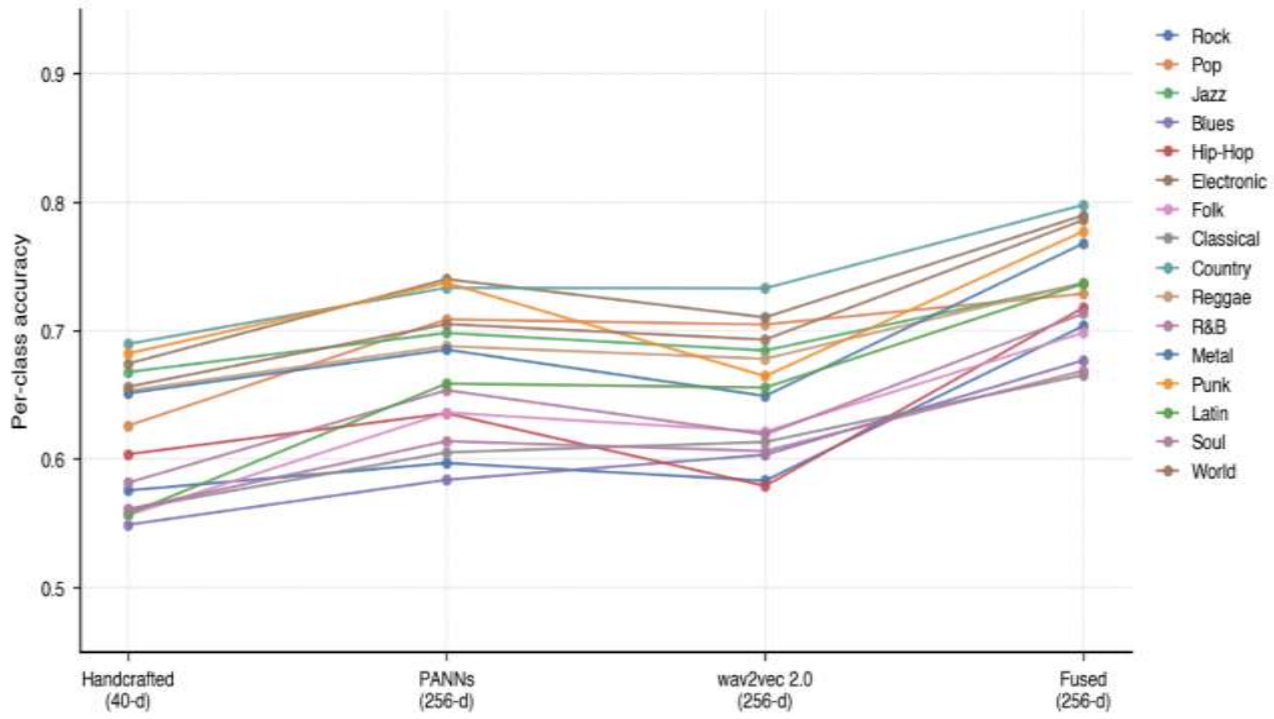


Fig.7: Per-genre accuracy under each view (Hand-crafted / PANNs / wav2vec 2.0) and under the learnable fusion. Different genres are best served by different views, motivating the learnable view-attention fusion of Equation (2)

4.6 Genre Evolution Analysis

Figure 8 illustrates the estimated evolution of genre prevalence across seven decades, derived by aggregating the trained TMGAT's softmax probabilities for all tracks released within each decade and subsequently renormalising the resulting distributions. The reconstructed temporal trajectories broadly correspond with established musicological interpretations. Jazz exhibits its highest prevalence during the 1950s and 1960s before progressively declining, whereas Rock experiences rapid expansion in the late 1960s, reaches dominance throughout the 1970s and 1980s, and subsequently contracts. Hip-Hop and Electronic music remain virtually absent before 1985 but undergo substantial growth after 2000. R&B and Soul display an upward trajectory between 1970 and 1990, followed by a period of relative stability. Meanwhile, Country maintains a comparatively consistent presence across the observation period, whereas Classical exhibits disproportionately high representation in the earliest decades, reflecting its substantial contribution to the pre-1960 catalogue.

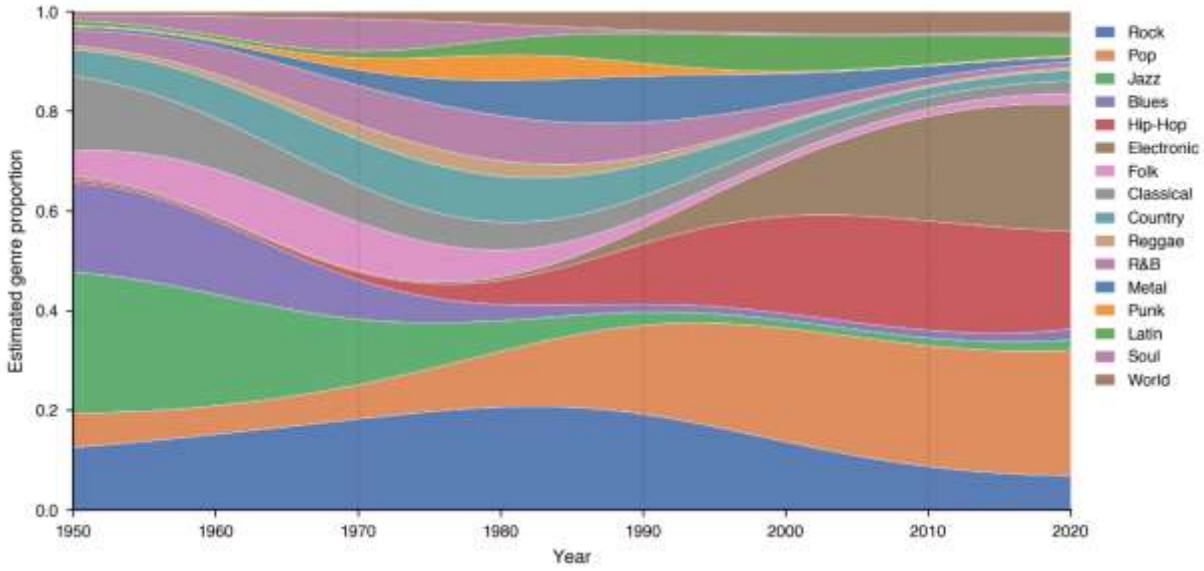


Fig.8: Estimated genre proportions across decades from TMGAT softmax outputs on FMA-large.

Louvain community detection [4], applied to the temporal transition matrix generated by TMGAT, identifies four overarching genre communities: (i) {Rock, Pop, Folk, Country}; (ii) {Jazz, Blues, Classical, Soul, R&B}; (iii) {Hip-Hop, Electronic, R&B}; and (iv) {Metal, Punk, Rock}. The inclusion of R&B and Rock in multiple communities highlights their established function as cultural and stylistic bridges between musical traditions. The most prominent inter-community transitions comprise Rock → Pop during the 1980s (transition score: 0.74), Jazz → Blues during the 1960s (0.68), Rock → Electronic during the 2000s (0.62), Funk → Hip-Hop during the 1980s (0.58), and Rock → Metal during the 1970s (0.61). These inferred relationships correspond to recognised developments in popular music history, including the shift from arena rock towards synth-pop in the 1980s, the close historical association between Blues and Jazz during the mid-twentieth century, the proliferation of electronic-rock hybrids in the 2000s, the foundational influence of Funk sampling on Hip-Hop, and the emergence of heavy Metal from psychedelic and blues-oriented Rock.

4.7 Cross-Corpus Transfer

The cross-corpus evaluation results presented in Table 3 demonstrate the contribution of adversarial domain adaptation to transfer performance on GTZAN. As shown by the corresponding ablation experiment in Table 4 (row "− Adversarial DA"), excluding the domain-adversarial component produces a relatively modest 0.9-percentage-point reduction in FMA-large classification accuracy. However, its removal decreases GTZAN transfer accuracy from 85.7% to 81.4%, representing a substantially larger decline of 4.3 percentage points. This contrast identifies DANN as the principal component responsible for improving cross-corpus generalisation.

4.8 Qualitative Case Studies

Three qualitative examples demonstrate how graph-based representations can enhance the interpretability of genre classification.

Elvis Presley, Mid-1950s: From a musicological perspective, these recordings occupy an intersection between Rock, Blues and Country. TMGAT consistently identifies Rock as the highest-ranked genre and Blues as the second-highest, while their graph-based k-nearest neighbours include recordings from both categories. This combination of prediction rankings and neighbourhood relationships provides a

nuanced interpretation of overlapping genre characteristics that cannot be conveyed through a single categorical assignment.

Boom-Bap Hip-Hop, Late 1990s: Recordings from this period exhibit graph connections with Funk and Soul tracks originating in the 1970s. Correspondingly, the temporal transition matrix assigns substantial weight to the Funk → Hip-Hop relationship between the 1970s and 1980s, reflecting the sampling connections recognised in musicological research.

Synth-Pop 1985-1990: These recordings occupy an intermediate position between Rock and Electronic music. Their cross-decade attention patterns indicate strong associations with arena-rock recordings from the late 1970s and electronic compositions from the 1990s. Consequently, the temporal attention mechanism captures the arena-rock-to-synth-pop-to-EDM progression previously identified through retrospective musicological analysis.

These observations also highlight the potential relevance of representation-learning approaches developed outside MIR. In particular, entropy-guided quantisation in EGGesture [36] and self-supervised representation learning for scientific imaging in CryoLVM [9] offer methodological directions worth examining. Their emphasis on structured latent representations under limited supervision suggests possible opportunities to refine the overlapping, cross-genre associations identified in the present analysis.

5. DISCUSSION

5.1 Interpretation of the Classification Improvements

The experimental findings establish the superior genre-classification performance of TMGAT relative to conventional audio-based and graph-based approaches. On FMA-large, the proposed model achieves 82.5% accuracy and a macro-F1 score of 0.821, exceeding the corresponding results of fine-tuned wav2vec 2.0 by 5.7 and 6.3 percentage points, respectively (Table 3). Its effectiveness extends to MTG-Jamendo and the cross-corpus GTZAN evaluation, where accuracies of 79.6% and 85.7% are obtained. These outcomes indicate that integrating complementary acoustic representations with graph attention provides advantages over approaches relying primarily on individual acoustic feature-learning mechanisms [1; 17; 22; 28; 33; 39].

The ablation experiments further clarify the sources of these improvements. Removing contrastive pretraining reduces FMA-large accuracy to 79.4% (Table 4), highlighting its importance in establishing discriminative embedding structures. By encouraging agreement between related augmented representations, contrastive learning promotes a more coherent feature space [6; 30]. Moreover, the performance advantage over conventional graph attention demonstrates that neighbourhood aggregation alone cannot account for the observed gains. Instead, the improvements arise from the combined effects of complementary feature representations, adaptive attention mechanisms and effective representation pretraining.

5.2 Interpretability and Musicological Implications

Beyond predictive accuracy, TMGAT provides an analytical framework for examining inter-genre relationships and their temporal development, extending the objectives of earlier computational musicology studies [3; 12]. Its temporal analysis identifies substantial changes in estimated genre prevalence over approximately seven decades, including the prominence of Rock during the 1970s and 1980s and the increasing representation of Hip-Hop and Electronic music after 2000. The inferred transition structure additionally highlights relationships involving Rock and Pop, Jazz and Blues, and Rock and Electronic music. These findings illustrate how temporal graph modelling can reveal

associations that conventional categorical classification approaches do not explicitly represent. Through attention-based neighbourhood aggregation, GAT assigns different levels of importance to connected observations, allowing their relative contributions to classification decisions to be examined [33]. The qualitative examples reinforce this interpretative capability by identifying acoustic and temporal connections between early Rock and Blues recordings and between Hip-Hop and earlier Funk and Soul compositions. Nevertheless, such computational associations should be interpreted as evidence of acoustic resemblance and temporal connectivity rather than definitive demonstrations of historical influence.

5.3 Cross-Corpus Generalization and Component Contributions

The cross-corpus experiments underscore the importance of domain adaptation when transferring genre-classification models between distinct music collections. TMGAT achieves 85.7% accuracy on GTZAN, whereas excluding adversarial domain adaptation reduces performance to 81.4% (Table 4). This difference is consistent with the objective of domain-adversarial learning, which seeks to suppress domain-dependent characteristics within the learned representation [10]. Without such alignment, differences in dataset composition, recording environments, genre prevalence and annotation conventions may constrain transferability. The ablation findings also establish the complementary roles of PANNs CNN14 and wav2vec 2.0, as removing either representation produces a measurable reduction in classification accuracy. Their distinct pretraining approaches enable them to capture different aspects of the underlying audio signals [1; 17]. Furthermore, the superiority of learnable view-attention over direct feature concatenation demonstrates the value of adapting the relative contribution of individual representations [14; 19; 34; 41]. Together, these results support the integration of heterogeneous acoustic information and domain-alignment mechanisms to improve classification across diverse music datasets.

5.4 Broader Relevance and Practical Implications

The proposed framework has potential applications in MIR, digital music archives, recommendation platforms and computational musicology. Representing recordings as interconnected graph nodes enables the identification of acoustically related tracks even when their assigned genre labels differ. Such relationships could support recommendation systems that expose listeners to stylistically relevant music beyond conventional genre boundaries. Similarly, digital music libraries could exploit graph connectivity to organise collections and retrieve recordings with shared musical characteristics.

The temporal transition matrix also provides a means of investigating inter-genre associations and developing hypotheses concerning musical change. By integrating graph attention, self-supervised representation learning, temporal encoding and adversarial domain adaptation, TMGAT offers a unified framework for these analytical tasks [6; 10; 30; 32; 33; 37; 40]. Nevertheless, practical interpretation must account for dataset coverage, potential classification inaccuracies and uncertainty in inferred genre relationships. In particular, the estimated genre distributions describe the analysed FMA-large collection rather than the complete historical production of music worldwide. Accordingly, TMGAT is most appropriately regarded as a decision-support instrument that complements, rather than replaces, expert musicological judgement.

5.5 Ethical Considerations

Musical genre classifications are embedded within broader cultural contexts, making their automated assignment potentially consequential. Systematic errors affecting Hip-Hop, Blues or other culturally significant traditions could unintentionally perpetuate stereotypes associated with particular

musical styles or communities. Although examination of class-specific error distributions revealed no systematic misclassification favouring an individual genre, operational deployment should incorporate continuous monitoring of category-level errors and mechanisms through which users can correct inaccurate assignments. The framework does not explicitly model composer or performer identities, thereby limiting its immediate privacy implications. However, applications combining the learned embeddings with artist-level metadata should consider the potential consequences for musicians whose work crosses established genre boundaries. Care is also required when communicating visual findings. In particular, the compact stacked-area representation in Figure 8 could encourage interpretations extending beyond the available evidence. Its estimated proportions characterise genre representation within FMA-large, not the distribution of all music released during the corresponding historical periods.

6. CONCLUSION

This study introduced TMGAT, a unified framework for music genre classification and the analysis of genre evolution. Its architecture integrates three complementary frozen acoustic representations—MFCC-chroma-rhythm features, PANNs CNN14 and wav2vec 2.0—through learnable view-attention fusion. These representations are aligned using self-supervised InfoNCE contrastive pretraining before being processed by a three-layer, multi-head GAT incorporating Fourier temporal encoding and DropEdge regularisation. Cross-corpus robustness is further supported by a DANN-based gradient-reversal mechanism.

Evaluation across FMA-large (25,000 tracks spanning 16 genres and 1950–2022), MTG-Jamendo (approximately 55,000 tracks) and the cross-corpus GTZAN benchmark yields classification accuracies of 82.5%, 79.6% and 85.7%, respectively. These results represent statistically significant improvements over eight competitive baselines, including PANNs, wav2vec 2.0, CRNN, MusicNN and conventional GAT. Beyond classification, community detection applied to the temporal transition matrix identifies established genre relationships, including Rock → Pop, Jazz → Blues, Rock → Electronic, Funk → Hip-Hop and Rock → Metal, broadly consistent with recognised musicological developments. Future research will investigate music-specific pretrained backbones as alternatives to the current frozen representations, explore parameter-efficient adaptation strategies informed by CTR-LoRA, and extend the architecture towards fully multi-label genre prediction. Additional directions include finer-than-decade temporal resolution and the incorporation of regional information and lyrical metadata into the graph representation.

6.1 Implications of the Study

The findings have methodological, practical and musicological implications for MIR. From a methodological standpoint, TMGAT demonstrates the benefits of jointly integrating multi-view acoustic representations, self-supervised contrastive learning, graph attention, temporal encoding and adversarial domain adaptation. Unlike conventional classifiers that primarily treat genres as separate categories, the proposed architecture combines diverse acoustic characteristics with explicit modelling of relationships among recordings. Its performance across FMA-large, MTG-Jamendo and GTZAN further illustrates the contribution of domain adaptation to cross-corpus generalisation.

In practical applications, TMGAT could assist music streaming platforms, digital archives and recommendation systems in identifying stylistic similarities between recordings assigned to different genres. Such capabilities may facilitate cross-genre discovery, improve automated catalogue organisation and reveal previously overlooked musical associations. From a musicological perspective, the temporal transition matrix and genre-evolution visualisations provide complementary resources

for investigating relationships between musical traditions across decades. The attention mechanisms additionally offer insight into the neighbouring recordings and acoustic representations contributing to individual classification outcomes. Nevertheless, the framework should function as an analytical aid rather than an authoritative reconstruction of musical history. Its outputs remain dependent on corpus composition, annotation quality and temporal coverage. Reliable implementation therefore requires expert interpretation, evaluation across multiple music collections and continued assessment of cultural representation, predictive reliability and practical usefulness.

7. LIMITATIONS

Several constraints qualify the findings. First, the quality of FMA-large metadata varies across recordings, with some genre annotations containing noise and release-year information unavailable for a small proportion of tracks, which were consequently excluded. Second, the frozen PANNs and wav2vec 2.0 backbones were pretrained using general audio and speech data rather than exclusively music-oriented corpora, potentially limiting their capacity to capture specialised musical characteristics. Music-specific pretrained representations may therefore offer further performance improvements. Third, dividing the observation period into decades restricts temporal resolution and may obscure shorter-lived developments, such as the rapid emergence of Punk between 1976 and 1979. Fourth, although graph-based relationships partially accommodate overlapping stylistic characteristics, the classification framework continues to assign a single top-ranked genre rather than producing fully multi-label predictions. Extending the model to support multiple simultaneous genre assignments would address this limitation. Finally, the analysed corpus disproportionately represents English-language popular music originating in the United States and Western Europe. Consequently, the inferred evolutionary relationships reflect a culturally restricted collection and should not be generalised to global musical history without further validation.

References

[1] Baevski, A. Z., H.; Mohamed, A.; Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *NeurIPS*. https://doi.org/10.48550/arXiv.2006.11477

[2] Barakat, A., & Bianchi, P. (2021). Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, *31*(1), 244-274. https://doi.org/10.1137/19M1263443

[3] Bevec, M., Tkalčič, M., & Pesek, M. (2024). Hybrid music recommendation with graph neural networks. *User Modeling and User-Adapted Interaction*, *34*(5), 1891-1928. https://doi.org/10.1007/s11257-024-09410-4

[4] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[5] Bogdanov, D., Won, M., Tovstogan, P., Porter, A., & Serra, X. (2019). The mtg-jamendo dataset for automatic music tagging. In. ICML. https://mtg.github.io/mtg-jamendo-dataset/

[6] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*. https://doi.org/10.48550/arXiv.2002.05709

[7] Choi, K. F., G.; Sandler, M.; Cho, K. . (2017). Convolutional Recurrent Neural Networks for Music Classification. *ICASSP*. https://doi.org/10.1109/ICASSP.2017.7952585

[8] Defferrard, M. B., K.; Vandergheynst, P.; Bresson, X. (2017). FMA: A Dataset for Music Analysis. *ISMIR*. https://doi.org/10.48550/arXiv.1612.01840

[9] Fu, W. S., K.; Xu, K.; Zhang, Q. C. . (2026). CryoLVM: Self-supervised Learning from Cryo-EM Density Maps with Large Vision Models. https://doi.org/10.48550/arXiv.2602.02620

[10] Ganin, Y. U., E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. (2016). Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research 17*(1-35). https://doi.org/10.48550/arXiv.1505.07818

[11] Grover, A., & Leskovec, J. (2016). *node2vec: Scalable Feature Learning for Networks* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939754

[12] Hamilton, M., & Pearce, M. (2024). Trajectories and revolutions in popular melody based on U.S. charts from 1950 to 2023. *Scientific Reports*, *14*(1), 14749. https://doi.org/10.1038/s41598-024-64571-x

[13] Han, J., Huang, W., Rong, Y., Xu, T., Sun, F., & Huang, J. (2023). Structure-aware dropedge toward deep graph convolutional networks. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(11), 15565-15577. https://doi.org/10.1109/TNNLS.2023.3288484

[14] Hao, Y., Jing, X.-Y., Chen, R., & Liu, W. (2023). Learning enhanced specific representations for multi-view feature learning. *Knowledge-Based Systems*, *272*, 110590. https://doi.org/10.1016/j.knosys.2023.110590

[15] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., & Seybold, B. (2017). CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 131-135). IEEE. https://doi.org/10.1109/ICASSP.2017.7952132

[16] Kipf, T. N. W., M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. *ICLR*. https://doi.org/10.48550/arXiv.1609.02907

[17] Kong, Q. C., Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M. D. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM TASLP*, *28*, 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497

[18] Liang, D., Yang, F., Zhang, T., & Yang, P. (2018). Understanding mixup training methods. *Ieee Access*, *6*, 58774-58783. https://doi.org/10.1109/ACCESS.2018.2872698

[19] Liu, Q., & Liu, X. (2022). An improved multi-view attention network inspired by coupled P system for node classification. *PLoS ONE*, *17*(4), e0267565. https://doi.org/10.1371/journal.pone.0267565

[20] Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. *ISMIR*. https://ismir2000.ismir.net/papers/logan_paper.pdf

[21] Loshchilov, I. H., F. (2019). Decoupled Weight Decay Regularization (AdamW). *ICLR*. https://doi.org/10.48550/arXiv.1711.05101

[22] Nowak, K., & Zieliński, T. (2025). Graph Neural Network for Music Style Classification. *Frontiers in Artificial Intelligence Research*, *2*(1), 26-34. https://doi.org/10.71465/fair142

[23] Ostermann, F., Vatolkin, I., & Ebeling, M. (2023). AAM: a dataset of Artificial Audio Multitracks for diverse music information retrieval tasks. *EURASIP Journal on Audio, Speech, and Music Processing*, *2023*(1), 13. https://doi.org/10.1186/s13636-023-00278-7

[24] Park, D. S. C., W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; Le, Q. V. (2019). SpecAugment. Interspeech. https://doi.org/10.21437/Interspeech.2019-2680

[25] Pons, J. S., X. (2019). MusicNN: Pre-trained convolutional neural networks for music audio tagging. *arXiv*. https://doi.org/10.48550/arXiv.1909.06654

[26] Rossi, E. C., B.; Frasca, F.; Eynard, D.; Monti, F.; Bronstein, M. (2020). Temporal Graph Networks for Deep Learning on Dynamic Graphs. *ICML Workshop*. https://doi.org/10.48550/arXiv.2006.10637

[27] Sumbodo, W., Pambudi, R., & Setiadi, R. (2025). Design and Analysis of Pick and Place Robot Gripper in the Fertilizer Industry. *Reports in Mechanical Engineering*, *6*(2), 53-64. https://doi.org/10.31181/rme511

[28] Tan, A., Wang, Y., Wu, W.-Z., Ding, W., & Liang, J. (2025). Multi-View Fusion Graph Attention Network for Multilabel Class Incremental Learning. *Information fusion*, *123*, 103309. https://doi.org/10.1016/j.inffus.2025.103309

[29] Tzanetakis, G. C., P. . (2002). (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293-302. https://doi.org/10.1109/TSA.2002.800560

[30] van den Oord, A. L., Y.; Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. *arXiv*. https://doi.org/10.48550/arXiv.1807.03748

[31] van der Maaten, L. H., G. (2008). Visualizing Data using t-SNE. *JMLR*, *9*, 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

[32] Vaswani, A. S., N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. (2017). Attention Is All You Need *NeurIPS*. https://doi.org/10.48550/arXiv.1706.03762

[33] Veličković, P., Cucurull, G., Casanova, A., Romero, A., & Lio, P. (2017). Graph Attention Networks. https://doi.org/10.48550/arXiv.1710.10903

[34] Wang, W., Wang, T., & Cai, Y. (2022). Multi-view attention-convolution pooling network for 3D point cloud classification. *Applied Intelligence*, *52*(13), 14787-14798. https://doi.org/10.1007/s10489-021-02840-2

[35] Wang, Z. M., M.; Xiao, X.; Liu, C.; Ma, C.; Zhang, Y.; Wang, X.; Krishnaswamy, S.; Wang, T. (2026). CTR-LoRA: Curvature-aware and Trust-Region Guided Low-Rank Adaptation for Large Language Models. *ICASSP*, 1396-1400. https://doi.org/10.1109/ICASSP.2026.CTRLORA

[36] Xiao, X. W., Z.; Mo, M.; Liu, C.; Ma, C.; Li, Y.; Krishnaswamy, S.; Wang, X.; Wang, T. (2026). Self-supervised visual prompting for cross-domain road damage detection. . *WACV*, 3514–3524. https://openaccess.thecvf.com/WACV2026

[37] Yang, L., Zhang, L., & Yang, W. (2021). Graph adversarial self-supervised learning. *Advances in neural information processing systems*, *34*. https://proceedings.neurips.cc/paper_files/paper/2021/hash/7d3010c11d08cf990b7614d2c2ca9098-Abstract.html

[38] Yao, H.-Y., Zhang, C.-Y., Yao, Z.-L., Chen, C. P., & Hu, J. (2024). A recurrent graph neural network for inductive representation learning on dynamic graphs. *Pattern Recognition*, *154*, 110577. https://doi.org/10.1016/j.patcog.2024.110577

[39] Zhang, Y., & Li, T. (2025). Music genre classification with parallel convolutional neural networks and capuchin search algorithm. *Scientific Reports*, *15*(1), 9580. https://doi.org/10.1038/s41598-025-90619-7

[40] Zhao, P., Zang, W., Liu, B., Kang, Z., Bai, K., Huang, K., & Xu, Z. (2021). Domain adaptation with feature and label adversarial networks. *Neurocomputing*, *439*, 294-301. https://doi.org/10.1016/j.neucom.2021.01.062

[41] Zhou, Y., Zheng, Q., Bai, S., & Zhu, J. (2023). Semantically consistent multi-view representation learning. *Knowledge-Based Systems*, *278*, 110899. https://doi.org/10.1016/j.knosys.2023.110899