时间异质图对比学习在多模态声事件分类中的应用

Yuanjian Chen* Yang Xiao† Jinjie Huang*

*Harbin University of Science and Technology
†The University of Melbourne

ABSTRACT

多模态声学事件分类在视听系统中起着关键作用。尽管结合音频和视觉信号可以提高识别效果，但在时间上对齐它们并减少跨模态噪声的影响仍然很困难。现有方法通常将音频和视频流分别处理，后期通过对比或互信息目标融合特征。近期的研究探索了多模态图学习，但大多数未能区分模内与模间的时间依赖性。为了解决这个问题，我们提出了时间异构图对比学习（THGCL）。我们的框架为每个事件构建一个时序图，其中音频和视频片段形成节点，它们的时序链接形成边。我们引入高斯过程来实现模内平滑，霍克斯过程来实现模间衰减，并通过对比学习捕捉细粒度的关系。在 AudioSet 上的实验表明，THGCL 实现了最先进的性能。

Index Terms— 异构图，多模态，对比学习，声事件分类

1. 介绍

多模态声学事件分类（AEC）[1, 2] 是智能音视频系统中的一个重要任务。它支持许多实际应用，包括安全监控、多媒体内容检索和人机交互 [3]。这些系统得益于结合音频和视觉信号来理解复杂环境。然而，在实际情况中，由于背景噪声、重叠声音或录音条件差等原因，音频信号往往不清楚。仅依赖音频可能会导致事件识别错误 [4]。为解决这一问题，物体运动、场景转换或唇部动作等视觉线索可以提供有价值的信息补充。因此，结合这两种模态可以提高识别性能。

†Corresponding author (email: yxiao9550@student.unimelb.edu.au)

尽管这种多模态方法显示出巨大的潜力，但也引入了新的挑战。其中一个关键难题在于建模音频和视觉输入之间的正确时间关系 [5, 6, 7] 。事件往往遵循严格的时序顺序，即使跨模态之间有微小的错位也会使模型感到困惑。因此，设计能够有效捕捉音视频数据间时间结构的系统是至关重要的。大多数现有的多模态方法都是先分别处理音频和视觉特征再将它们结合在一起。通常情况下，每种模式都由专门的神经网络进行编码，并且它的特征会在后期通过拼接 [8, 9, 10] 的方式融合。为了减少跨模态噪声，许多研究人员还引入了额外的学习目标，比如对比损失或互信息最大化 [4, 11] 。例如，跨模态师生框架 [12] 通过鼓励音频和视觉信号之间的协议来学习更稳健的嵌入表示。其他方法如 XDC [13] 和演变损失 [14] 已经证明了结合单模态和多模态预训练任务以改善表示学习的好处。

超越特征融合，基于图的学习为建模多模态关系提供了一条有前景的路径 [15]。尽管在 AEC 中仍不常见，这类方法已经在其他领域取得了成功。例如，图形已被用于连接图像-文本对 [16]，改进文本到语音中的说话风格 [17]，以及在社交媒体中建模用户和视频的关系 [18]。受此进展的启发，TMac [19] 将图学习应用于声学事件，通过将音频和视频片段表示为节点及其时间链接作为边来实现。然而，大多数基于图的方法仍然同等对待模内和跨模态关系，忽略了它们的时间差异。这通常会导致对齐不佳和表现力有限。

为了解决这些限制，我们提出了一种新的框架，称为时间异构图对比学习（THGCL）。我们的方法通过基于其时间一致性分配不同权重给模内和模间交互来显式建模节点之间的时间关系。我们采用高斯过程

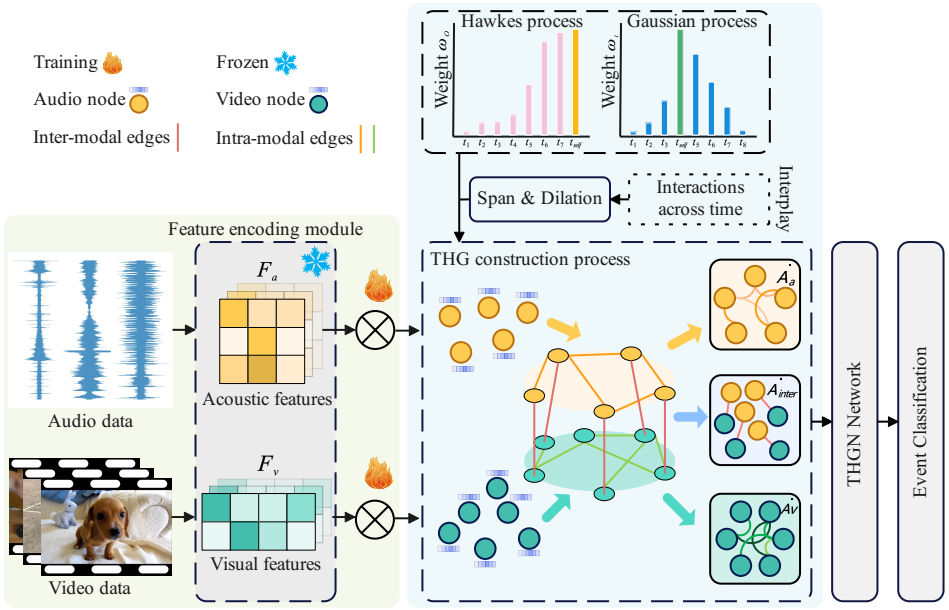

Fig. 1. 基于时间异构图的对比学习（THGCL）的整体框架

捕捉每个模态内的平滑性，并使用 Hawkes 过程模型跨模态的衰减依赖关系。此外，我们引入了一种对比学习目标以更好地减少跨模态噪声。通过这种设计，THGCL 提高了多模态表示的鲁棒性和表达能力。我们在 AudioSet 上评估了 THGCL。结果表明，我们的方法达到了最先进的性能。

2. 我们的方法

我们提出了一种新的框架，称为 THGCL，如图 1 所示。THGCL 构建了一个时空异构图，该图建模了跨模式的时序和语义依赖关系，用于声学事件分类。

2.1. 特征编码模块

我们首先提取音频和视觉特征以构建框架的基础。对于音频，我们首先将片段剪辑到 960 毫秒的段，并将这些段转换为 96×64 对数-梅尔频谱图。该频谱图被传递给 VGGish [20]，生成 128 维的音频特征。对于视频，每个片段被分割成非重叠的 250 毫秒块。每个块通过预训练的 S3D 网络 [21, 22] 处理，产生 1024 维的视频特征。

为了对齐维度，我们应用线性变换以生成嵌入 E_a 和 E_v，这两个嵌入的维度为 d。这些对齐后的嵌入随后作为图构建的输入。

2.2. 时间异质图的构建

我们将每个视听输入表示为一个时间异构图（THG）[19, 23, 24]。这里我们提供了一个图形构建过程的全面描述 G。我们首先将声学和视觉嵌入分别划分为 P_a 和 P_v 段。节点集 G 由声学节点 $\mathbb{V}^a = \{v_1, ..., v_{P_a}\}$ 和视觉节点 $\mathbb{V}^v = \{v_1, ..., v_{P_v}\}$ 组成，而边集则包括音频内部、视觉内部和跨模态的边，对应的邻接矩阵分别为 A_a,A_v 和 A_{inter}。定义图中边的跨模态交互由两个参数控制：膨胀率和时间跨度。对于声学节点而言，音频特征的细粒度要求精确的时间对齐。对于视觉节点，其交互受到空间分辨率和运动的影响，这允许较粗略的时间对齐。对于跨模态的边，必须严格遵循时间一致性，并且只有当时间戳匹配时才存在这些边。一旦构建了时间异构图，所得表示就与原始音视频数据保持一致。然后应用时间权重来优化边。

在**对于模内关系**中，我们使用高斯过程 [25, 26] 以确保时间上更接近的节点获得更高的权重。音频和视频的加权邻接矩阵定义为 $\bar{A}_a$ 和 $\bar{A}_v$，模态的时间权重为 m：

$$s_m^{i,j} = \exp\left(-\frac{||P_m^i - P_m^j||^2}{2 \times (P_m^{\max} - P_m^{\min} + 1)^2}\right), m \in \{a, v\}, \quad (1)$$

在这里，$P^{\max}$ 和 $P^{\min}$ 表示当前节点邻域内最大和最小的段索引。这些时间权重反映了嵌入空间中关系的

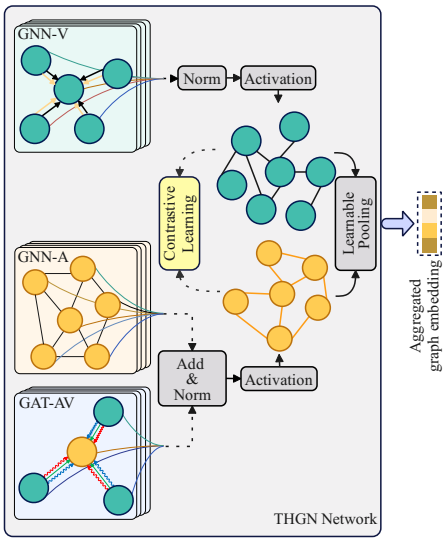

Fig. 2. THGN 网络的架构。

平滑度，在这个空间中，附近的片段在语义和特征上倾向于更加相似。

对于模态间关系，音频和视频的表示经常不同，有必要建模过去交互的衰减效果。我们应用霍克斯过程 [27] 来为这些边赋予权重，这将最近的交互赋予更强的影响。跨模式邻接定义为 $\bar{A}_{inter}$ 并带有时间权重：

$$s_{inter}^{i,j} = \sigma\left(\frac{\log \xi}{\log(1-\xi)} + \frac{P_a^{\max'} - P_v^i + 1}{P_a^{\max'} - P_a^{\min'} + 1}\right)/\tau, \tag{2}$$

在这里，$\xi \sim U(0,1)$ 引入了随机性，τ 控制平滑度，而 $\sigma(\cdot)$ 是 sigmoid 函数。权重反映了时间上更接近的节点具有更强的影响，而遥远事件逐渐失去其影响力。通过结合这些策略，我们获得加权邻接矩阵 $\bar{A}_a, \bar{A}_v, \bar{A}_{inter}$，它在模态之间保持了时间一致性。然后将这些矩阵用作图神经网络的输入。实际上，视听输入以批次形式进行处理，这允许并行训练并提高可扩展性。

2.3. 模型训练方案

为了学习跨模态的节点表示和边权重，我们提出了一种时间异质图网络（THGN）。目标是减少可能不支持声学分类的视觉特征的影响，其次，通过图聚合过滤出无关的声学细节。

给定嵌入集 $E = \{E_a, E_v\}$ 和加权邻接矩阵集 $\bar{A} = \{\bar{A}_a, \bar{A}_v, \bar{A}_{inter}\}$，GNN 编码器可以写成 $X^{out} = \Gamma(E, \bar{A}, \Psi)$，其中 Ψ 表示可学习的参数。第 l 层定义为：

$$X^l = \rho(\bar{A}X^{l-1}\Psi^{l-1}), l = 1, ..., \mathbb{L}, \tag{3}$$

以 $X^0 = E$ 作为输入，$X^{out} = X^{\mathbb{L}}$ 作为最终输出，并且 $\rho(\cdot)$ 是一种非线性激活函数，如 ReLU。

THGN 聚合了模内和模间信息，同时尊重时间权重。如图 2 所示，它使用四层处理音频和视觉输入，然后将视频信息整合到音频节点中。这确保模型专注于声学事件的同时受益于视觉线索。该网络包括四个时间图层、一个对比模块和一个可学习的池化层。音频节点 (X_a^l) 由 GNN-A 更新，视频节点 (X_v^l) 由 GNN-V 更新，并通过 GAT-AV 实现跨模态传输。

为了进一步提高表示并减少跨模态的噪声影响，我们添加了一个对比模块。与标准多模态对比学习不同，我们的方法在片段级别上使用异构图进行工作。我们设计了一项自我判别任务，要求相同目标在跨模态中的相似度高于同一模态内不同目标的相似度。受 SimCLR [28] 的启发，我们将从视频到音频的损失相应定义。

$$L_{v\to a} = -\frac{1}{\mathbb{B}}\sum_{i\in\mathbb{B}} \log \frac{\exp\left(\cos\left(X_a^{(i)}, X_v^{(i)}\right)/\mathfrak{t}\right)}{\sum_{j\in\mathbb{B}, j\neq i}\exp\left(\cos\left(X_a^{(i)}, X_v^{(j)}\right)/\mathfrak{t}\right)}, \tag{4}$$

其中 $\mathbb{B}$ 表示训练批次，$X_a^{(i)}$ 和 $X_v^{(i)}$ 是第 i 个样本的图嵌入，$\cos(\cdot)$ 是余弦相似度，而 $\mathfrak{t}$ 是温度参数。对称损失 $L_{a\to v}$ 从音频到视频被定义，最终的对比损失 L_{CL} 是 $L_{a\to v}$ 和 $L_{v\to a}$ 的和。

然后我们将它连接到分类任务。由于 THGCL 是为声事件分类设计的，我们应用一个可学习的池化函数 [19] 来聚合输出 X_a^{out} 和 X_v^{out} 以形成图级别的嵌入 X^G。此嵌入被传递到一个分类头部，该头部使用焦点损失 [29] 进行训练以解决类别不平衡问题。THGCL 的整体训练目标随后被表述为分类损失和对比损失的加权组合：

$$L = \omega_{FL}L_{FL} + \omega_{CL}L_{CL}, \tag{5}$$

在我们的实验中 $\omega_{FL} = 1.0$ 和 $\omega_{CL} = 0.1$。这个联合目标平衡了准确的事件分类与鲁棒的跨模态表示学习。

3. 实验

3.1. 实验设置

3.1.1. 数据集

我们的实验是在 AudioSet[30] 上进行的。每个音频片段的持续时间固定为 10 秒。我们通过选择 33 个其评分者置信度分数落在范围 $[0.7, 1.0]$ 内的声音事件类别来构建一个高置信子集，从而得到了 $82,410$ 个音视频训练样本。我们在原始评估集合上评估我们的方法，该集合包含 $85,487$ 个测试片段 [19]，以确保公平的比较。

3.1.2. 实现和度量

在特征编码模块中，我们将转换维度 d 设置为 128。对于时间异构图的构建，音频、视频和跨模态节点之间的时间跨度分别设置为 6、4 和 3，而同一模式内的时序扩张则设置为 3 和 4。在训练过程中，我们采用了初始学习率为 0.005 的 Adam 优化器。最大迭代次数限制为 5000，并启用了提前停止功能。THGN 网络的隐藏通道大小为 512。性能使用平均精度均值 (mAP) 和 ROC 曲线下的面积 (AUC) 进行评估。我们的代码发布在 [1] 这个仓库中。

3.2. 总体比较

表 1 显示，提出的 THGCL 仅使用 4.8M 参数就达到了 57.4%的最佳 mAP 和 0.948 的 AUC，显示出准确性和效率。表格 1 底部非基于图形的方法表现不佳的结果突显了利用图结构来建模复杂的时态关系对于声事件分类的重要性。如 TMac 和 VAED 这样的异构图模型捕捉到了结构性关系但忽略了时间对齐和衰减，限制了它们的提升。波形模型例如 Wave-Logmel 以及大型 ResNet-1D 变体在噪声敏感性和低效性方面挣扎。因此，它们表现不佳。相比之下，THGCL 统一了高斯和 Hawkes 加权与对比学习，捕捉到跨模态和多模态的时间依赖关系并抑制了噪声，从而达到了超越所有基线的最佳结果。

[1] https://github.com/visionchan/THGCL.git

Table 1. 音频事件检测在 Audioset 中的可比性研究。“⋆” 表示基于图的方法。

Model	mAP (%)	AUC	#Params (M)
THGCL (Ours)	57.4	0.948	4.8
TMac ⋆ [19]	55.1	0.937	4.3
VAED ⋆	51.6	0.919	2.1
PaSST-S	49.0	0.900	87.0
ASiT [31]	48.5	–	85.0
Audio-MAE (local)	48.2	–	86.0
ATST-clip [32]	47.8	–	86.0
MaskSpec	47.3	–	86.0
LHGNN ⋆ [33]	46.6	–	31.0
Conformer-based [34]	44.4	–	88.0
HGCN ⋆	44.2	0.885	42.4
Wave-Logmel	44.1	–	81.0
AST	44.0	–	88.0
SSL graph ⋆	43.9	–	0.2
Wav2vec2-audio	42.6	0.880	94.9
VATT	39.7	–	87.0
ResNet-1D-both	38.0	0.891	81.2
ConvNeXt-femto [35]	37.9	–	5.0
R(2+1)D-video	36.0	0.810	33.4
ResNet-1D-audio	35.9	0.900	40.4

Table 2. 不同时间类型的表现对比

ID	Temporal type	mAP (%)	AUC
1	w/ Gau. & Haw.	57.4	0.948
2	both Haw.	55.0	0.942
3	both Gau.	53.5	0.893

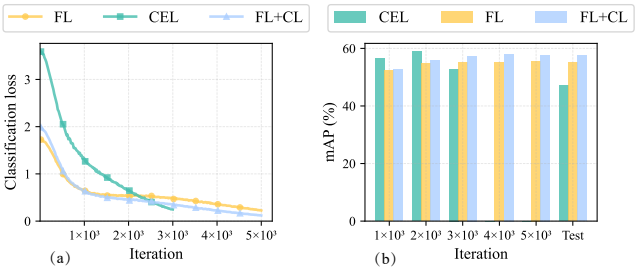

Fig. 3. 比较不同损失函数对声学事件分类的影响。(a) 训练损失与迭代次数的关系。(b) 不同阶段性能的变化。

3.3. 关系建模策略的消融研究

为了评估时间图构建对模型判别能力的有效性，我们首先在表 2 中比较了替代的时间关系建模策略。具体来说，ID-1 表示我们在第 2.2 节中提到的配置；ID-2 复制了 TMac 设置；ID-3 将高斯过程 (Gau.) 应用于所有时间边。结果显示 ID-1 达到了最佳性能。在

单模态内使用 Gau. 是有效的，因为声学事件表现出短期平稳性且相邻节点具有相似的表示。跨模态采用霍克斯过程 (Haw.) 更好地表征了跨模态激发和对齐，从而保持了时间一致性并增强了整体判别性能。

3.4. 不同损失函数的消融研究

我们还通过探索不同的损失函数来探讨自监督对比学习的贡献。figure 3(a) 显示交叉熵损失 (CEL) 收敛最慢且保持在较高的损失水平；焦距损失 (FL) 最初下降较快，但很快达到平台期；我们的损失 L (FL+CL) 实现了快速初始下降和持续改进，达到了最低且最平稳的最终损失。figure 3(b) 表明 FL 在早期 mAP 上优于 CEL，而 FL+CL 全程领先并保持稳定的中期到晚期优势。这表明加入对比损失使模型能够学习更强的跨模态表示，并减少噪声污染的影响，从而提高泛化能力和缓解过拟合。总体而言，联合损失在收敛速度、最终性能和稳定性方面都超过了分类损失。

4. 结论

我们提出了一种时间异质图对比学习（THGCL）方法来解决声事件分类中的跨模态时间不对齐和背景噪声干扰问题。THGCL 构建了一个时间异质图，其中同模态边使用高斯权重以保持平滑连贯性，而异模态边则使用 Hawkes 权重以实现同步。随后，时间异质图网络（THGN）将对比学习与嵌入聚合相结合，以增强跨模态一致性、扩大类别边界并抑制噪声段落。在 AudioSet 上的实验表明 THGCL 超越了强大的基线方法，并且其模块是互补的。

5. REFERENCES

[1] Huriye Atilgan, Stephen M Town, Katherine C Wood, Gareth P Jones, Ross K Maddox, Adrian KC Lee, and Jennifer K Bizley, "Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding," Neuron, vol. 97, no. 3, pp. 640–655, 2018.

[2] Yang Xiao and Rohan Kumar Das, "Wilddesed: an llm-powered dataset for wild domestic environment sound event detection system," in Proc. DCASE, 2024.

[3] Ibrahim Ghafir, Vaclav Prenosil, Jakub Svoboda, and Mohammad Hammoudeh, "A survey on network security monitoring systems," in Proc. IEEE FiCloudW, 2016, pp. 77–82.

[4] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song, "Active contrastive learning of audio-visual video representations," in Proc. ICLR, 2021.

[5] Stéphane Dupont and Juergen Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Transactions on Multimedia, vol. 2, no. 3, pp. 141–151, 2000.

[6] Han Yin, Jisheng Bai, Yang Xiao, Hui Wang, Siqi Zheng, Yafeng Chen, Rohan Kumar Das, Chong Deng, and Jianfeng Chen, "Exploring text-queried sound event detection with audio source separation," in Proc. ICASSP. IEEE, 2025, pp. 1–5.

[7] Yuanjian Chen, Yang Xiao, Han Yin, Yadong Guan, and Xubo Liu, "Noise-robust sound event detection and counting via language-queried sound separation," arXiv preprint:2508.07176, 2025.

[8] Hong Liu, Wanlu Xu, and Bing Yang, "Audio-visual speech recognition using a two-step feature fusion strategy," in Proc. ICPR, 2021, pp. 1896–1903.

[9] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj, "Avff: Audio-visual feature fusion for video deepfake detection," in Proc. CVPR, 2024, pp. 27102–27112.

[10] Yang Xiao, Han Yin, Jisheng Bai, and Rohan Kumar Das, "Mixstyle based domain generalization for sound event detection with heterogeneous training data," arXiv preprint:2407.03654, 2024.

[11] Aaqib Saeed, David Grangier, and Neil Zeghidour, "Contrastive learning of general-purpose audio representations," in Proc. ICASSP. IEEE, 2021, pp. 3875–3879.

[12] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba, "Ambient sound provides supervision for visual learning," in Proc. ECCV. Springer, 2016, pp. 801–816.

[13] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran, "Self-supervised learning by cross-modal audio-video clustering," Advances in neural information processing systems, vol. 33, pp. 9758–9770, 2020.

[14] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo, "Evolving losses for unsupervised video representation learning," in Proc. CVPR, 2020, pp. 133–142.

[15] Xihong Yang, Yue Liu, Sihang Zhou, Siwei Wang, Wenxuan Tu, Qun Zheng, Xinwang Liu, Liming Fang, and En Zhu, "Cluster-guided contrastive graph clustering network," in Proc. AAAI, 2023, vol. 37, pp. 10834–10842.

[16] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo, "A novel graph-based multi-modal fusion encoder for neural machine translation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3025–3035.

[17] Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su, "Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling," in Proc. ICASSP, 2022, pp. 7917–7921.

[18] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua, "Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video," in Proc. ACM MM, 2019, pp. 1437–1445.

[19] Meng Liu, Ke Liang, Dayu Hu, Hao Yu, Yue Liu, Lingyuan Meng, Wenxuan Tu, Sihang Zhou, and Xinwang Liu, "Tmac: Temporal multi-modal graph learning for acoustic event classification," in Proc. ACM MM, 2023, pp. 3365–3374.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in Proc. ICASSP. IEEE, 2017, pp. 131–135.

[21] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in Proc. ECCV, 2018, pp. 305–321.

[22] Tengda Han, Weidi Xie, and Andrew Zisserman, "Self-supervised co-training for video representation learning," Advances in neural information processing systems, vol. 33, pp. 5679–5690, 2020.

[23] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla, "Heterogeneous graph neural network," in Proc. ACM SIGKDD, 2019, pp. 793–803.

[24] Shirian Amir, Krishna Somandepalli, Victor Sanchez Silva, and Tanaya Guha, "Visually-aware acoustic event detection using heterogeneous graphs," in Proc. INTERSPEECH. ISCA, 2022, pp. 2428–2432.

[25] Jinyuan Fang, Shangsong Liang, Zaiqiao Meng, and Qiang Zhang, "Gaussian process with graph convolutional kernel for relational learning," in Proc. ACM SIGKDD, 2021, pp. 353–363.

[26] Jiazheng Li, Chunhui Zhang, and Chuxu Zhang, "Heterogeneous temporal graph neural network explainer," in Proc. CIKM, 2023, pp. 1298–1307.

[27] Alan G Hawkes, "Point spectra of some mutually exciting point processes," Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 33, no. 3, pp. 438–443, 1971.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in Proc. ICML. PmLR, 2020, pp. 1597–1607.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in Proc. ICCV, 2017, pp. 2980–2988.

[30] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in Proc. ICASSP. IEEE, 2017, pp. 776–780.

[31] Sara Atito Ali Ahmed, Muhammad Awais, Wenwu Wang, Mark D Plumbley, and Josef Kittler, "Asit: Local-global audio spectrogram vision transformer for event classification," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 3684–3693, 2024.

[32] Xian Li, Nian Shao, and Xiaofei Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 1336–1351, 2024.

[33] Shubhr Singh, Emmanouil Benetos, Huy Phan, and Dan Stowell, "Lhgnn: Local-higher order graph neural networks for audio classification and tagging," in Proc. ICASSP. IEEE, 2025, pp. 1–5.

[34] Sangeeta Srivastava, Yun Wang, Andros Tjandra, Anurag Kumar, Chunxi Liu, Kritika Singh, and Yatharth Saraf, "Conformer-based self-supervised learning for non-speech audio tasks," in Proc. ICASSP. IEEE, 2022, pp. 8862–8866.

[35] Thomas Pellegrini, Ismail Khalfaoui-Hassani, Etienne Labbé, and Timothée Masquelier, "Adapting a convnext model to audio classification on audioset," in Proc. INTERSPEECH. ISCA, 2023, pp. 4169–4173.