Audio Multi-View Spoofing Detection Framework Based on Audio-Text-Emotion Correlations

24年 TIFS（开源）

1. 目标：音频，伪造检测

2. 核心思想：用音频多视角特征（音频 - 文本 - 情感）的相关性捕获视角内和视角间线索，实现音频欺骗检测，而不依赖特定伪造痕迹

3. 方法：AMSDF：前端特征提取器+多视角相关性测量网络+后端合成分类器

4. 前端特征提取器：①文本特征：使用预训练的ASR模型，即 wav2vec2.0XLS-R300M，并整个框架内进行微调；②音频特征：音频编码器由 Sinc 层和各种残差块组成，对欺骗线索进行建模；③情感特征：SER 模型，在IEMOCAP数据集上进行预训练，并微调

5. 多视角相关性测量网络：

   ①视角内图注意力机制 IGAM，聚合同一视角内的节点

   ②异构图融合模块 HGFM，测量视角间节点的相关性

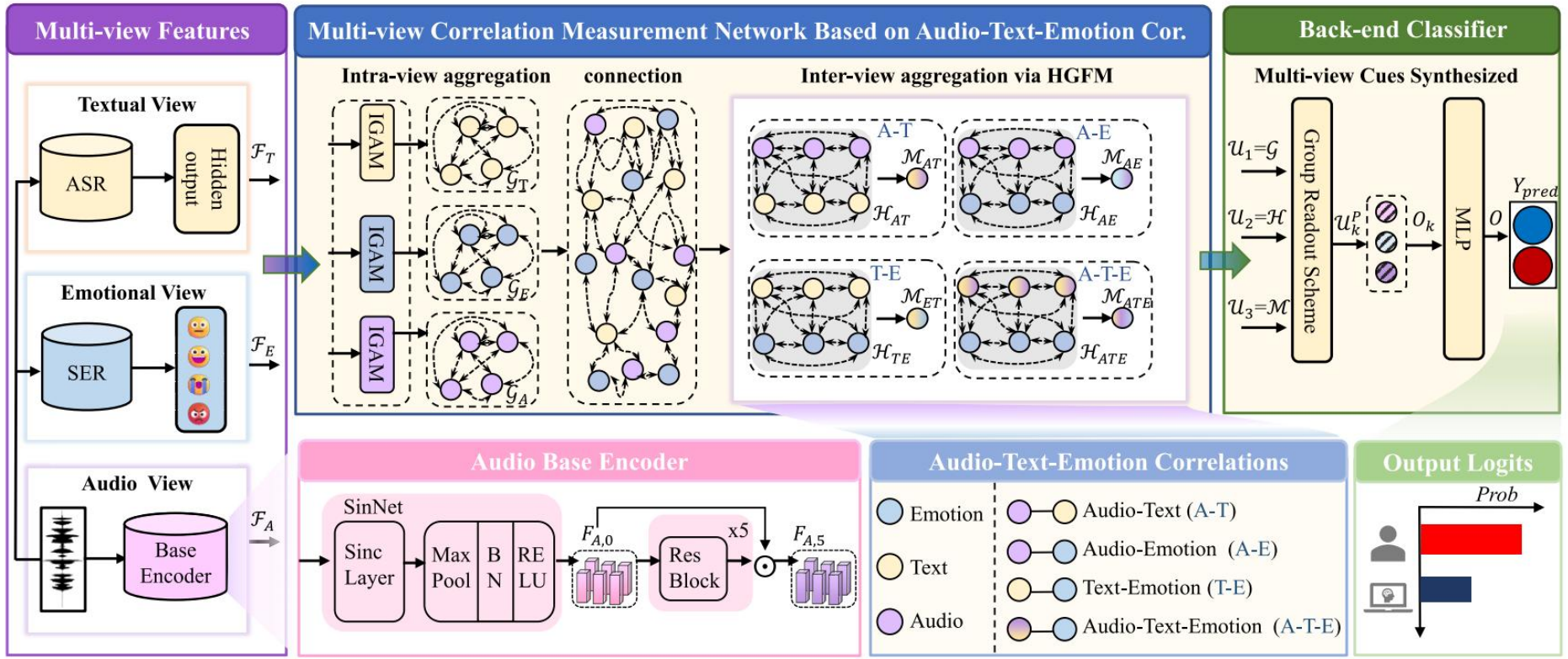
6. 后端合成分类器：

   ①表示不同的欺骗线索分组

   ②基于组的读出方案（GRS）

   ③获得组级输出

   ④输入多层感知器（MLP）以生成二进制预测



Audio Multi-View Spoofing Detection Framework Based on Audio-Text-Emotion Correlations

24年 TIFS（开源）

7. 视角内图注意力机制（IGAM）：

①构建初始图：同一视图中，每个节点与其他所有节点相连；

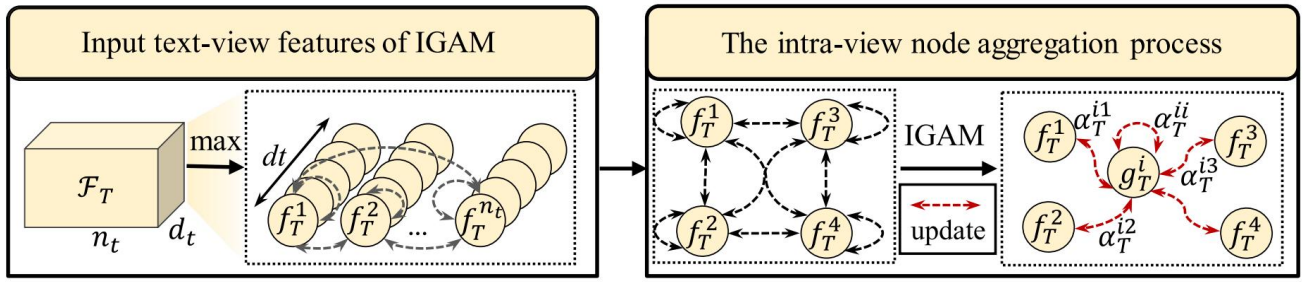
②计算注意力权重：表示节点之间关系的强度，W是可学习映射，用于调整视角内关系权重； $e_v^{ij} = W_v \cdot \left(f_v^{(i)} \odot f_v^{(j)} \right)$，

③归一化注意力权重：使用 softmax 函数对注意力权重进行归一化； $\alpha_v^{ij} = \mathrm{softmax}_j\left(e_v^{ij}\right) = \frac{\exp\left(e_v^{ij}/\tau\right)}{\sum_{l \in \mathcal{N}(f_v^{(i)})} \exp\left(e_v^{il}/\tau\right)}$，

④计算聚合信息：通过其连接节点的加权和计算； $f_v^{(i)\prime} = \sum_{j \in \mathcal{N}(f_v^{(i)})} \alpha_v^{ij} f_v^{(j)}$，

$$g_v^{(i)} = \mathrm{SELU}\left(W_{agg}\left(f_v^{(i)\prime}\right) + W_{res}\left(f_v^{(i)}\right)\right),$$

⑤更新节点：agg用于转换包含视角内聚合信息的节点，res用于投影原始节点，得到更新后的同构图



8. 异构图融合模块（HGFM）：

①构建异构图：- 连接节点构建初始异构图 $h_{AT}^{(i)} = \begin{cases} W_A^h g_A^{(i)}, & \text{if } i \in N(\mathcal{G}_A), \\ W_T^h g_T^{(i)}, & \text{if } i \in N(\mathcal{G}_T), \end{cases}$

- 测量节点间关系强度（类似于7.2，得到 e_{AT}^{ij}）
- 计算平均信息 $m_{AT} = \frac{1}{n_c}\sum_{i=1}^{n_c} h_{AT}^{(i)}$，
- 调整节点间相关性 $(e_{AT}^{ij})' = \begin{cases} \lambda_A^i e_{AT}^{ij}, & \text{if } i,j \in N(\mathcal{G}_A), \\ \lambda_T^i e_{AT}^{ij}, & \text{if } i,j \in N(\mathcal{G}_T), \\ \lambda_{AT}^i e_{AT}^{ij}, & \text{otherwise}, \end{cases}$
- 计算归一化权重并聚合节点

归一化权重 α_{AT}^{ij} 并 $h_{AT}^{(i)\prime} = \sum_{j \in \mathcal{N}(h_{AT}^{(i)})} \alpha_{AT}^{ij} h_{AT}^{(j)}$，

- 更新节点完成异构图构建（类似于7.5）

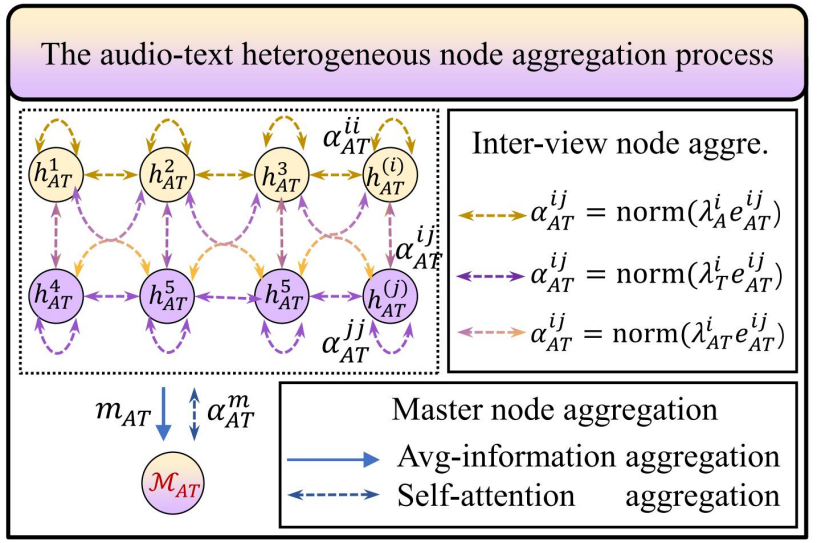
②获取主节点：

- 计算主节点重要性权重

$\alpha_{AT}^m = \mathrm{softmax}_j\left(W_{AT}^m \cdot \left(h_{AT}^{(j)} \odot m_{AT}\right)\right)$

$m'_{AT} = \sum_{j \in N(\mathcal{H}_{AT})} \alpha_{AT}^m h_{AT}^{(j)}$，

- 更新主节点（类似于7.5）



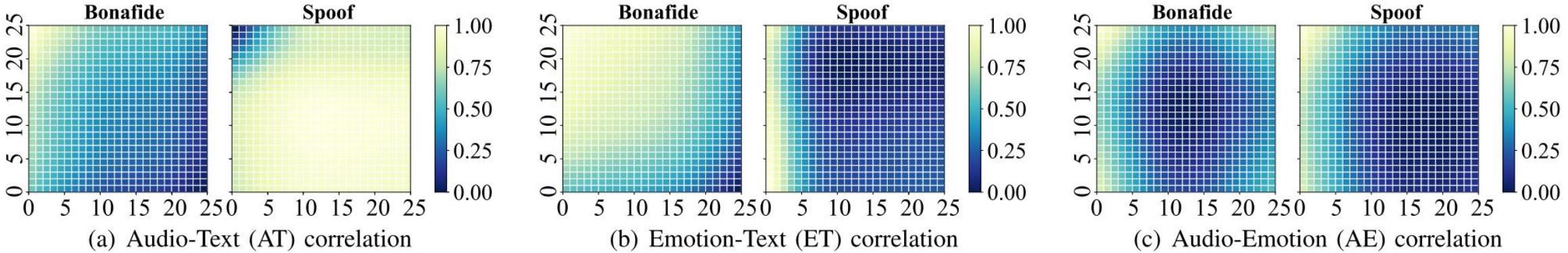
Audio Multi-View Spoofing Detection Framework Based on Audio-Text-Emotion Correlations

24年 TIFS（开源）

9. 异构图中边缘权重的可视化展示，体现了视角间相关性，热图中颜色越亮，表示相应的视角间节点之间的相关性越强：

①欺骗模式（Spoof）中 AT 相关性强于真实模式（Bonafine），AE 相关性差异弱于 ET 相关性差异；

②文本视图特征有助于区分真实和欺骗音频



(a) Audio-Text (AT) correlation (b) Emotion-Text (ET) correlation (c) Audio-Emotion (AE) correlation

在 ASVS2019LA 评估数据集上所获得的检测性能比较结果

Methods	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	EER	MT-DCF
C-GMM	**0.00**	**0.04**	0.14	15.16	**0.08**	4.74	26.15	10.85	1.26	**0.00**	19.62	3.81	**0.04**	9.57	0.2366
L-GMM	12.86	0.37	**0.00**	18.97	0.12	4.92	9.57	1.22	2.22	6.31	7.71	3.58	13.94	8.09	0.2116
E-DNN	0.22	3.42	0.11	0.45	0.33	0.19	0.19	0.42	0.38	0.75	20.51	17.20	6.24	7.14	0.1691
R-ST	0.34	0.53	0.08	0.38	0.29	0.33	0.30	0.29	0.27	0.83	1.71	3.11	0.73	1.07	0.0339
A-ST	0.47	0.41	**0.00**	0.79	0.16	0.67	0.12	0.16	0.51	0.65	1.28	2.65	0.67	0.82	0.0272
A-ST-L	0.61	0.20	**0.00**	1.06	0.16	0.77	0.20	0.06	0.61	0.65	1.91	3.03	0.69	0.99	0.0317
TSSDN	1.43	0.75	0.02	1.75	0.06	0.18	0.06	0.11	2.05	1.25	6.01	1.14	1.44	1.62	0.0474
SSLAS	0.06	0.06	0.02	0.40	0.10	**0.14**	**0.00**	0.06	0.24	0.06	0.37	0.84	0.35	0.25	0.0071
PSDL	0.69	1.04	0.76	1.79	0.76	0.76	0.68	1.69	2.28	0.71	0.81	1.85	4.78	1.67	0.0537
AMSDF†	0.03	0.05	0.03	0.52	0.38	0.23	0.02	0.05	0.33	0.12	0.30	0.47	0.41	0.31	0.0097
AMSDF	0.02	**0.04**	0.02	**0.23**	0.12	0.16	0.02	**0.02**	**0.16**	0.06	**0.26**	**0.29**	0.31	**0.16**	**0.0055**