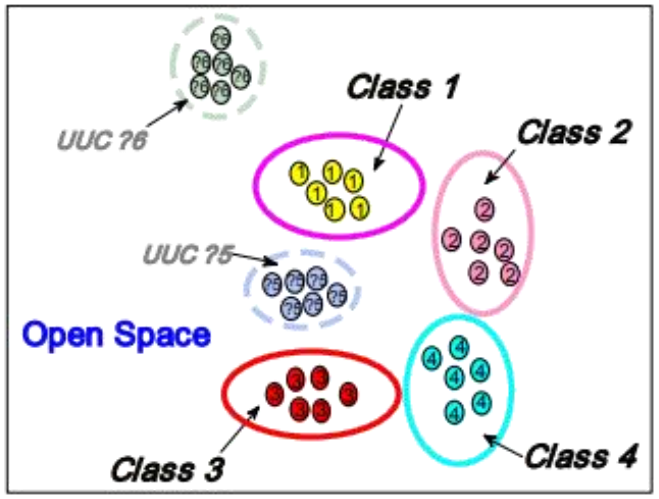
南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

PRISM: PROGRESSIVE ROBUST LEARNING FOR OPEN-WORLD CONTINUAL CATEGORY DISCOVERY

ICLR 2026

Background

open-set recognition (OSR)



(c) Open set recognition/classification problem.

novel-category discovery (NCD)

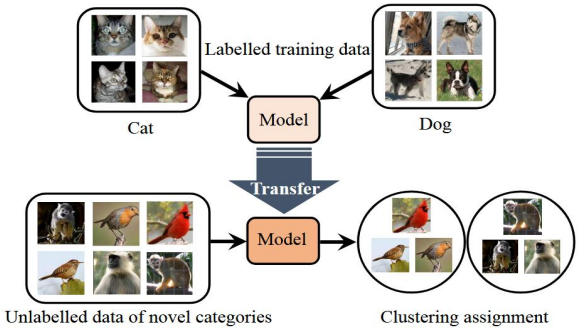


Figure 1. Learning to discover novel visual categories via deep transfer clustering. We first train a model with labelled images (e.g., cat and dog). The model is then applied to images of unlabelled novel categories (e.g., bird and monkey), which transfers the knowledge learned from the labelled images to the unlabelled images. With such transferred knowledge, our model can then simultaneously learn a feature representation and the clustering assignment for the unlabelled images of novel categories.

Generalized Category Discovery（GCD）

属性	OSR（开放集识别）	NCD（新类发现）	GCD（广义新类发现）
核心任务目标	区分“已知类”与“未知类”，未知类拒识（不分类）	在无标签数据中聚类出不同新类结构	在混合数据中，分类已知类+聚类新类
是否做新类分类	✗不需要（未知类标记为“未知”）	☑聚类发现新类的类别结构	☑聚类发现新类的类别结构
数据构成	有标签：仅已知类 无标签：测试阶段输入（已知类 + 未知类）	有标签：仅已知类 无标签：仅新类（无已知类）	有标签：仅已知类 无标签：混合（已知类 + 新类）
数据阶段特征	训练仅用已知类， 测试引入未知类	训练用已知类， 测试全为无标签新类	训练：已知类标签 + 混合无标签数据 推理：预测混合无标签数据的类别

Motivation

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

➢ 不足之处

1、**NCD/GCD（新类发现/广义新类发现）：**

1）依赖静态数据集，要求同时获取标记与未标记数据；

2）未考虑开放环境的动态性，无法适配现实中的数据流场景。

2、**Continual Category Discovery(CCD，持续新类发现)**

1）各阶段数据来自单一固定域，与开放环境的实际情况不符，样本多为多源/跨域数据。

2）新类别会伴随着域偏移（如设备、风格、光照变化）与已有类别同步出现。

开放世界场景下，连续新类发现+域偏移

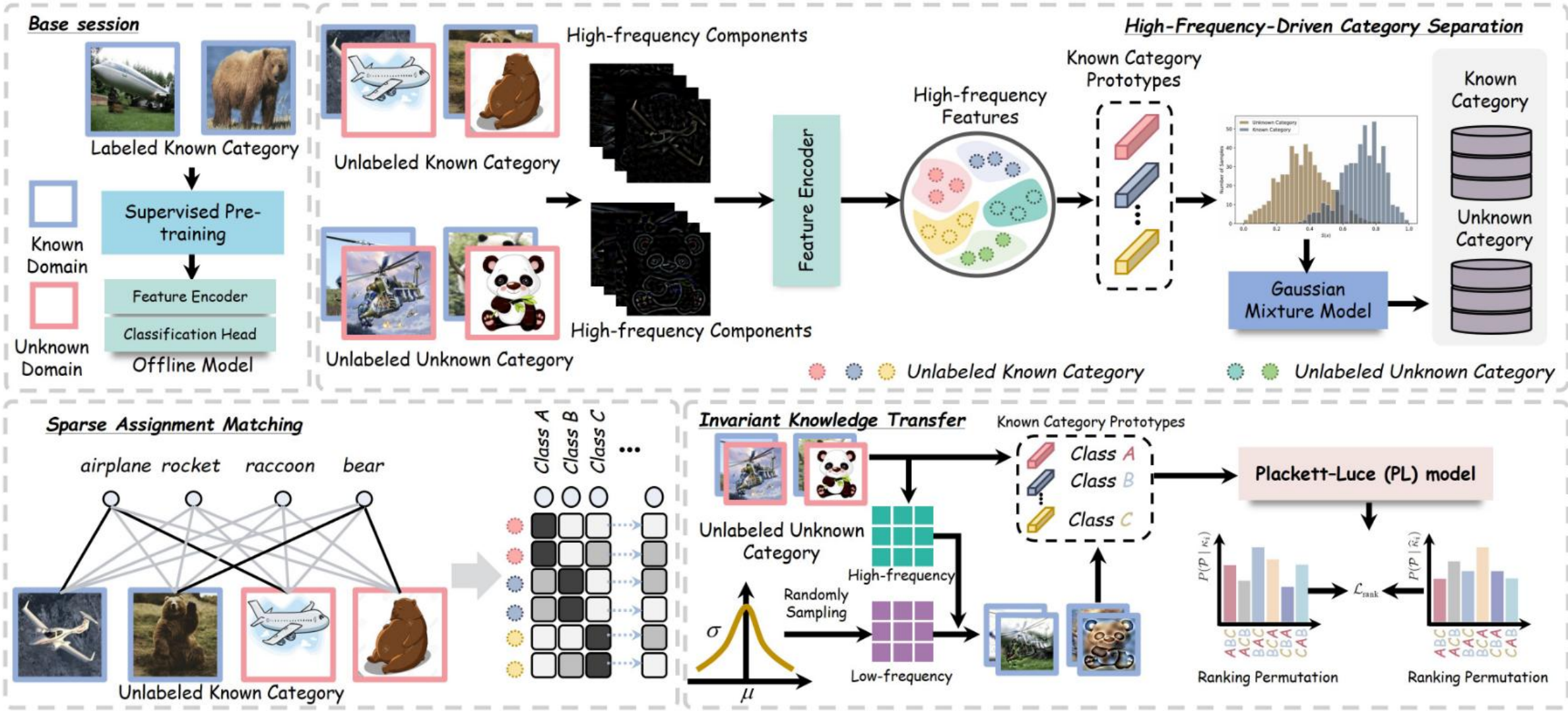
➢ 核心目标

1、分布偏移下保留已知类识别的能力

2、动态非平衡数据流持续发现新类别

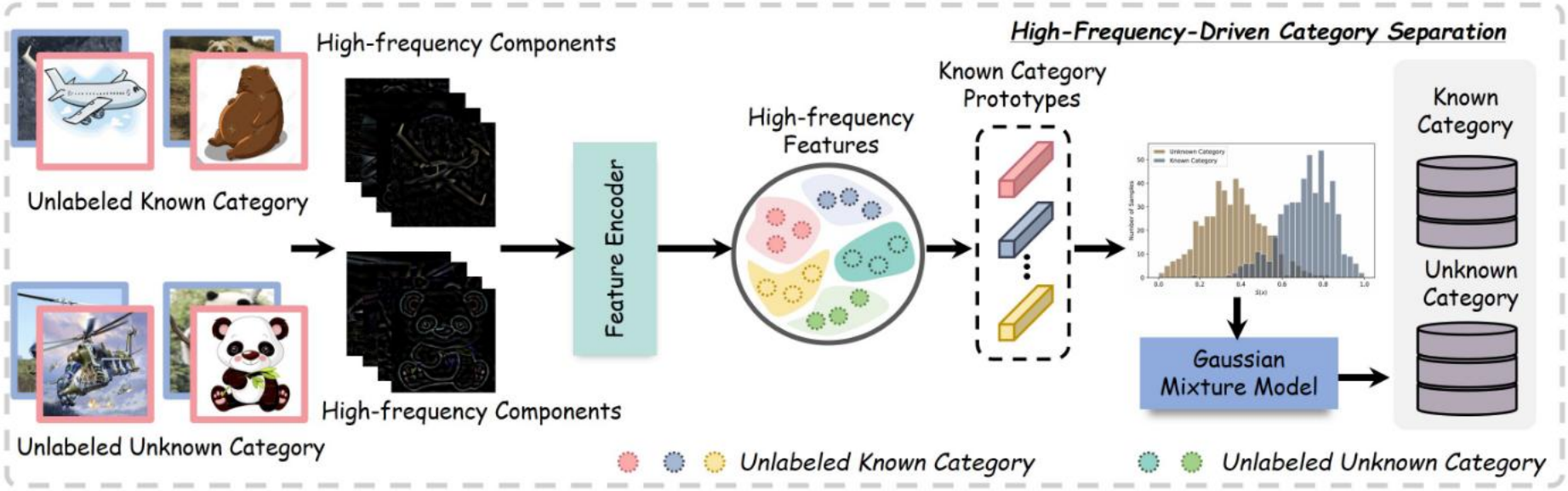
3、规避传统的域适应方法的局限性（标签空间重叠假设、负迁移风险、缺乏未知类发现指导）

Methods


Base session
Labeled Known Category
Known Domain
Supervised Pre-training
Feature Encoder
Classification Head
Unknown Domain
Offline Model
Unlabeled Known Category
Unlabeled Unknown Category
High-frequency Components
High-frequency Components
Feature Encoder
High-frequency Features
Known Category Prototypes
High-Frequency-Driven Category Separation
Gaussian Mixture Model
Known Category
Unknown Category
Unlabeled Known Category
Unlabeled Unknown Category
Sparse Assignment Matching
airplane rocket raccoon bear
Class A
Class B
Class C
Unlabeled Known Category
Invariant Knowledge Transfer
Known Category Prototypes
Class A
Class B
Class C
Unlabeled Unknown Category
High-frequency
Randomly Sampling
Low-frequency
Plackett-Luce (PL) model
Ranking Permutation
Ranking Permutation

Methods

➢ **高频驱动的类别分离（High-frequency-driven Category Separation，HCS）**



高频分量倾向于捕获域不变的全局语义（如结构），而低频分量则编码域相关的细节（如风格）。

(a) Image (b) Low-frequency (c) High-frequency

Methods

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- **高频驱动的类别分离（High-frequency-driven Category Separation，HCS）**

1、使用离散傅里叶变换到频域

$$\mathcal{F}(x_i)(u,v,c)=\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}x_i(h,w,c)\,e^{-j2\pi\left(\frac{hu}{H}+\frac{wv}{W}\right)},$$

2、二元掩码（分离高低频）

$$M_{u,v}=\begin{cases}1, & \text{if } \max(|u-\frac{H}{2}|,|v-\frac{W}{2}|)\le r\cdot\frac{\min(H,W)}{2},\\ 0, & \text{otherwise,}\end{cases}$$

3、高低频率分量获取

$$\mathcal{F}^l(x_i)=M\odot\mathcal{F}(x_i),\quad \mathcal{F}^h(x_i)=(I-M)\odot\mathcal{F}(x_i),$$

4、逆DFT恢复低高频图像的空间表示

$$x_i^l=\mathcal{F}^{-1}(\mathcal{F}^l(x_i)),x_i^h=\mathcal{F}^{-1}(\mathcal{F}^h(x_i)).$$

5、基于特征，计算分数

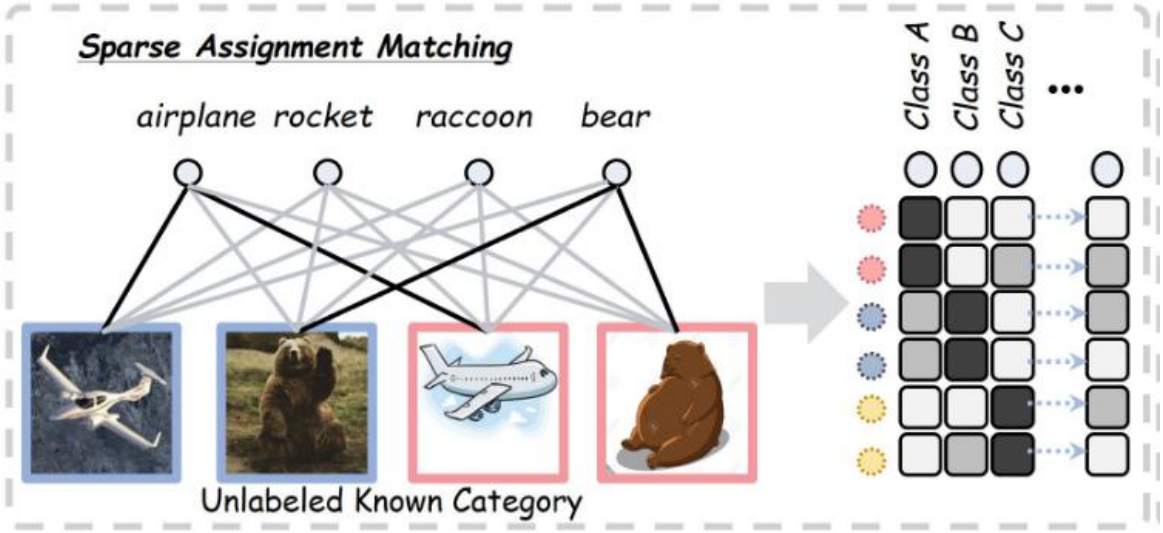
$$S(x)=\nu\left(\max_c\frac{f(x^h)\cdot e_c}{\|f(x^h)\|\cdot\|e_c\|}\right)$$

6、GMM区分已知和未知类

$$\mathcal{D}^u_{t,\text{kno}}=\{x\mid x\in\mathcal{D}^u_t\ \wedge\ \pi(x)\ge 0.5\},\quad \mathcal{D}^u_{t,\text{unk}}=\{x\mid x\in\mathcal{D}^u_t\ \wedge\ \pi(x)<0.5\}.$$

Methods

- **稀疏对齐匹配（Sparse Assignment Matching，SAM）**



Optimal Transport（最优运输）：自动学习“无标签已知样本”与“已知类原型”之间的最佳匹配，从而在跨域情况下恢复正确的类别对应关系。

- **不足之处**

1）直接用线性规划解 OT：计算成本极高，难以实际落地；
2）熵正则化 OT：虽然提升了效率，但会得到过于密集的传输计划，导致样本与原型的匹配结果不准确。

提出稀疏对齐匹配方案，通过引入L_2范数近邻项

Methods

- **稀疏对齐匹配（Sparse Assignment Matching， SAM）**

优化目标

$$\min_{\gamma\in\Delta}\ \ell(\gamma)=\sum_{i=1}^{N_{t,\mathrm{kno}}}\sum_{j=1}^{\mathcal{C}^{t-1}}\left[\gamma_{ij}\,C_{ij}\ +\ \frac{\varepsilon}{2}\left(\gamma_{ij}-\gamma_{ij}^{(l)}\right)^2\right],$$

约束集合

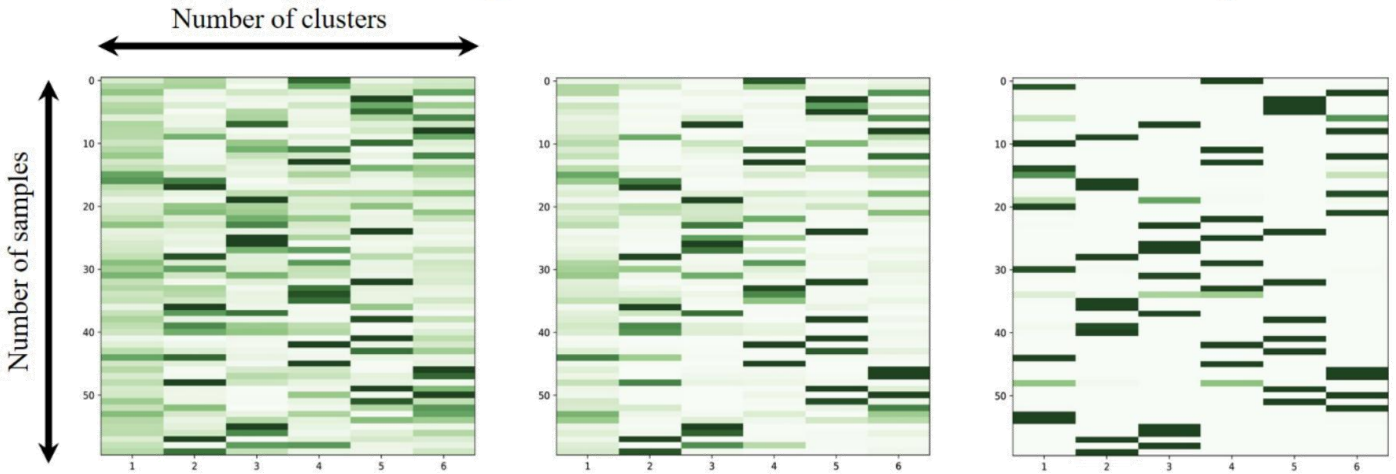
$$\Delta=\left\{\sum_{j=1}^{\mathcal{C}^{t-1}}\gamma_{ij}=\hat{a}_i=1,\sum_{i=1}^{N^{t,\mathrm{kno}}}\gamma_{ij}=\hat{b}_j=\frac{N_{t,\mathrm{kno}}}{\mathcal{C}^{t-1}},\gamma_{ij}\geq 0\right\}$$

通过对偶化（Dual form）避免直接处理约束

$$\max_{\psi,\varphi}\ \sum_{i=1}^{N_{t,\mathrm{kno}}}\psi_i\hat{a}_i\ +\ \sum_{j=1}^{\mathcal{C}^{t-1}}\varphi_j\hat{b}_j\ -\ \frac{\varepsilon}{2}\sum_{i=1}^{N_{t,\mathrm{kno}}}\sum_{j=1}^{\mathcal{C}^{t-1}}\left[\frac{\psi_i+\varphi_j-\tilde{C}_{ij}}{\varepsilon}\right]_+^2,$$

更新运输计划

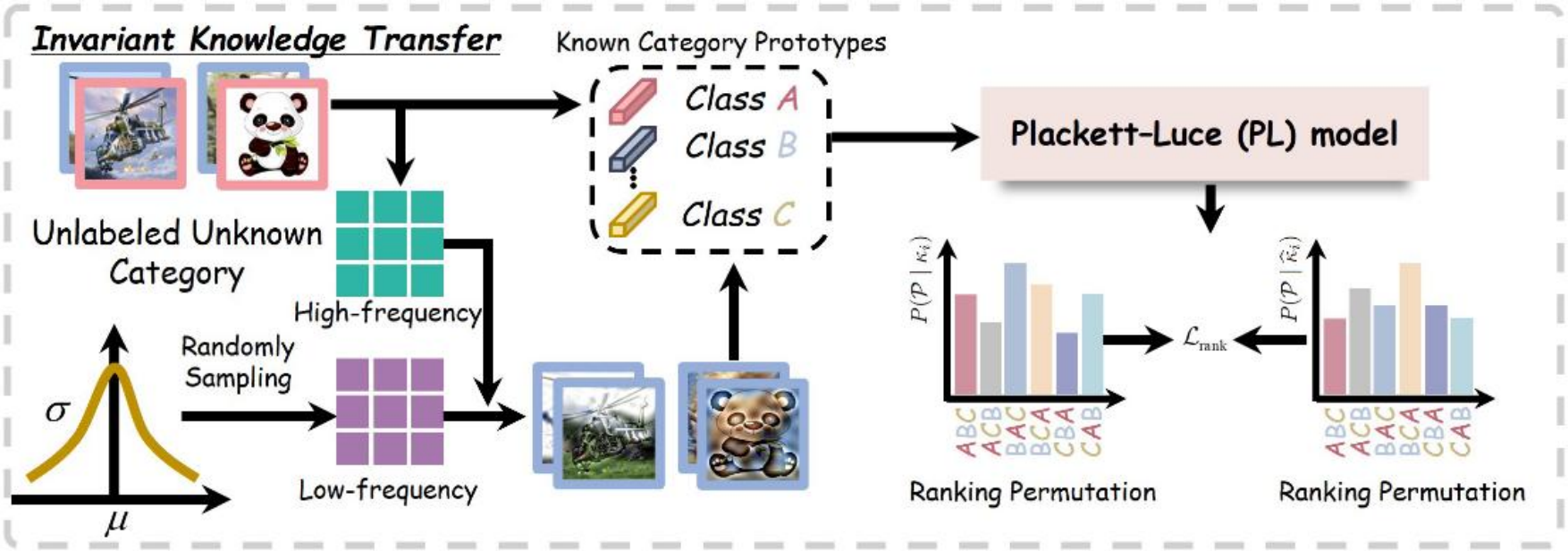
$$\gamma_{ij}^{(l+1)}=\max\left\{0,\ (\psi_i+\varphi_j+\varepsilon\gamma_{ij}^{(l)}-C_{ij})/\varepsilon\right\}.$$



(a) Original sinkhorn algorithm (b) ℓ_2-Norm optimal transport (c) Our proposed SAM method

Methods

- **不变知识转移（Invariant Knowledge Transfer，IKT）**



- 现有的不足

1）依赖语义关联的知识转移在域偏差下会被风格因素扭曲

2）现有方法缺乏对“域不变类别关联”的建模

通过建模并约束未知样本与已知原型之间的跨域关系保持一致，从而抵抗域偏移干扰，让模型依赖真正稳定的语义结构完成类别发现。

Methods

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- **不变知识转移（Invariant Knowledge Transfer，IKT）**

1、低频统计（style）的建模与扰动

1）用均值和反差表示低频风格

$$\mu(\mathcal{F}_i^l) = \frac{1}{HW}\sum_{u,v}\mathcal{F}_i^l(u,v,c), \quad \sigma(\mathcal{F}_i^l) = \frac{1}{HW}\sum_{u,v}[\mathcal{F}_i^l(u,v,c) - \mu(\mathcal{F}_i^l)]^2.$$

2）估计样本间统计量分布

$$\Sigma_\mu^2(\mathcal{F}_i^l) = \frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}}\left[\mu(\mathcal{F}_i^l) - \mathbb{E}[\mu(\mathcal{F}_i^l)]\right]^2, \quad \Sigma_\sigma^2(\mathcal{F}_i^l) = \frac{1}{N_{t-1}}\sum_{i=1}^{N_{t-1}}\left[\sigma(\mathcal{F}_i^l) - \mathbb{E}[\sigma(\mathcal{F}_i^l)]\right]^2.$$

3）高斯分布中采样受扰的低频数据（模拟域偏移）

$$\hat{\mu}(\mathcal{F}_i^l) = \mu(\mathcal{F}_i^l) + \epsilon_\mu \Sigma_\mu(\mathcal{F}_i^l), \quad \epsilon_\mu \sim \mathcal{N}(0,1), \quad \hat{\sigma}(\mathcal{F}_i^l) = \sigma(\mathcal{F}_i^l) + \epsilon_\sigma \Sigma_\sigma(\mathcal{F}_i^l), \quad \epsilon_\sigma \sim \mathcal{N}(0,1).$$

2、构建转换的低频特征

1）将未知样本低频替换为扰动后的低频

$$\widehat{\mathcal{F}}_{i,\text{unk}}^l = \hat{\sigma}(\mathcal{F}_i^l) \cdot \frac{\mathcal{F}_{i,\text{unk}}^l - \mu(\mathcal{F}_{i,\text{unk}}^l)}{\sigma(\mathcal{F}_{i,\text{unk}}^l)} + \hat{\mu}(\mathcal{F}_i^l).$$

3、未知样本与已知原型的关系建模

1）获得两种视图的特征

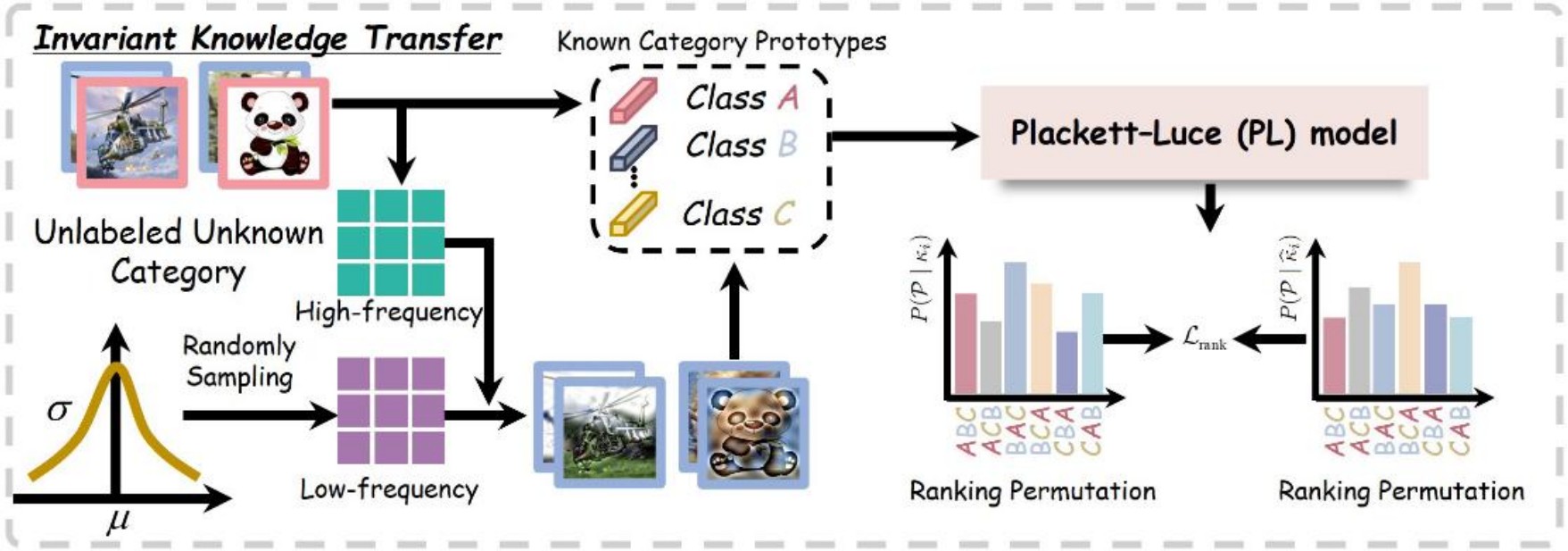
$$z_{i,\text{unk}}^t = f(x_{i,\text{unk}}^t), \widehat{z}_{i,\text{unk}}^t = f(\widehat{x}_{i,\text{unk}}^t).$$

2）计算未知样本与每个已知类原型的相似度

$$\kappa_{i,c} = \exp\left(\cos\left(z_{i,\text{unk}}^t, e_c^{t-1}\right)\right), \widehat{\kappa}_{i,c} = \exp\left(\cos\left(\widehat{z}_{i,\text{unk}}^t, e_c^{t-1}\right)\right).$$

Methods

- **不变知识转移（Invariant Knowledge Transfer，IKT）**



4、Plackett–Luce model：用全排序建模关系

$$P(\xi \mid \kappa_i) = \prod_{k=1}^{C^{t-1}} \frac{\kappa_{i,\xi(k)}}{\sum_{k'=k}^{C^{t-1}} \kappa_{i,\xi(k')}}.$$

5、强制两视图的排序分布一致（不变知识）

$$\mathcal{L}_{\text{rank}} = \frac{1}{N_{t,\text{unk}}} \sum_{i=1}^{N_{t,\text{unk}}} \ell_{\text{KL}}(P(\mathcal{P} \mid \kappa_i),\, P(\mathcal{P} \mid \widehat{\kappa}_i)),$$

通过随机扰动未知样本的风格构造不同视图，再用 Plackett–Luce 排序模型捕获未知样本与已知类原型的全局关系，并强制两视图的排序分布一致，从而让模型学习到跨域稳定的语义关联。

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

➢ 数据集

1、**SSB-C**：扩展了语义偏移基准（SSB），包含 9 种破坏类型（5 个严重程度等级）与 3 个细粒度数据集。SSB作为已知类，SSB-C作为未知类。

2、**DomainNet**：含6个多样化领域的大规模数据集，涵盖数百个类别且领域差距显著。Real作为已知类，其他类别作为未知类。

➢ 实现细节

采用ViT-B/16作为骨干网络，每个阶段仅微调最后一个 Transformer 块：采用 SGD 优化器训练 30 个 epoch，批次大小设为 128；初始学习率为 0.1，通过余弦退火衰减至 1×10^{-4}，权重衰减固定为 5×10^{-5}

➢ 评估方案

持续聚类准确率（continual clustering accuracy，cACC）

$$\mathrm{cACC}_t = \frac{1}{t}\sum_{k=1}^{t}\mathrm{ACC}_k$$

ACC定义

$$\mathrm{ACC} = \frac{1}{|\mathcal{D}_t^u|}\sum_{i=1}^{|\mathcal{D}_t^u|}\mathbb{I}\{y_i = g^*(\hat{y}_i)\}$$

g*：最优置换映射（将预测聚类匹配到对应的真实类别）

Experiments

Table 1: Clustering performance on DomainNet benchmark. We use Real as the known domain and each of the remaining domains as the unknown domain. We report the average All / Old / New accuracy across all stages for both domains.

Methods	Real → Painting						Real → Sketch						Real → Quickdraw						Real → Clipart						Real → Infograph					
	Real			Painting			Real			Sketch			Real			Quickdraw			Real			Clipart			Real			Infograph		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD	51.3	67.2	45.4	27.4	26.7	28.1	52.3	65.7	41.7	9.2	14.5	10.1	38.7	56.2	29.6	5.0	4.7	5.8	46.7	65.7	40.1	14.5	21.2	10.1	39.8	55.3	32.4	8.1	9.8	6.4
SimGCD	48.4	63.9	41.3	22.6	22.4	23.5	48.5	60.2	36.5	7.2	11.3	9.2	32.4	50.3	23.5	4.2	4.0	5.1	40.2	58.8	33.5	10.3	18.8	8.2	33.6	49.2	27.8	6.7	7.8	5.2
SPTNet	49.8	64.5	42.5	24.1	23.5	24.3	49.9	62.3	37.8	7.9	11.7	9.6	34.8	52.6	24.8	4.9	4.6	5.5	43.1	60.3	35.9	11.6	19.3	8.9	35.9	51.4	29.8	7.2	8.0	5.9
RLCD	50.8	66.2	44.1	25.5	24.6	25.8	51.2	64.8	40.1	8.4	12.1	10.0	36.1	54.0	25.7	4.8	4.7	5.3	45.2	62.1	36.9	13.5	20.9	9.8	37.1	53.2	32.5	8.4	8.9	6.8
G&M	47.1	62.3	41.2	26.3	25.5	26.2	50.9	63.4	42.3	10.9	15.1	10.5	34.1	50.2	27.3	4.3	4.1	5.2	40.3	61.1	34.2	11.4	19.2	8.8	32.4	50.1	27.6	7.5	9.2	5.5
Happy	50.6	66.5	44.7	28.0	27.1	28.9	52.0	65.0	41.2	11.2	15.6	10.7	35.6	51.4	28.9	4.6	4.5	5.2	45.6	62.4	37.1	12.0	19.6	9.0	34.2	50.5	28.0	7.9	9.4	5.6
PA-CGCD	55.4	70.3	48.1	30.1	30.8	30.2	55.1	70.7	46.6	12.3	16.1	11.2	43.6	60.4	34.2	5.1	5.0	6.0	52.2	70.3	44.6	17.8	24.5	12.3	45.2	61.3	38.1	9.0	11.8	7.1
DEAN	56.0	71.7	47.9	32.8	34.4	31.5	56.7	71.5	47.6	12.9	16.8	11.2	44.0	61.0	35.1	5.3	5.1	6.2	55.1	**72.7**	47.5	20.3	26.7	15.0	46.7	62.3	40.8	9.5	12.5	7.9
PromptCCD	56.5	71.2	50.3	31.5	32.1	31.2	57.4	**73.6**	48.6	13.4	17.7	12.1	45.2	62.3	36.7	5.8	5.1	6.5	54.1	71.2	46.7	19.8	26.1	14.4	47.1	63.1	40.2	9.2	12.2	7.8
VB-CGCD	57.3	71.0	52.4	32.4	33.6	32.5	56.9	73.1	48.8	13.9	18.1	12.9	47.1	62.1	38.1	6.0	4.9	6.8	55.4	72.0	47.5	19.6	25.8	14.2	48.3	63.9	41.9	9.4	12.4	8.0
PRISM	**60.9**	**74.1**	**55.1**	**39.2**	**39.0**	**38.2**	**60.1**	73.4	**51.0**	**16.9**	**20.1**	**15.9**	**54.0**	**74.0**	**49.2**	**7.1**	**6.5**	**7.4**	**58.0**	72.3	**51.2**	**24.0**	**30.4**	**19.1**	**60.1**	**73.8**	**53.1**	**10.9**	**14.1**	**9.8**

Table 2: Clustering performance on SSB-C benchmarks. Each dataset contains both Original and Corrupted settings, and we report the average All / Old / New accuracy across all stages for both domains.

Methods	CUB-C						Stanford Cars-C						FGVC-Aircraft-C					
	Original			Corrupted			Original			Corrupted			Original			Corrupted		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
GCD	29.4	47.7	23.4	26.8	45.9	20.1	26.4	56.1	21.5	22.3	43.1	11.2	27.7	33.6	24.9	28.8	41.4	28.8
SimGCD	26.6	44.5	21.0	23.4	42.4	17.7	23.1	52.5	18.9	19.3	39.7	9.8	25.4	30.1	22.1	25.2	38.1	25.8
SPTNet	27.8	45.2	22.0	25.1	44.2	18.1	24.9	55.0	20.3	21.1	41.6	9.9	26.1	31.2	23.3	26.9	39.5	26.7
RLCD	29.1	46.8	23.8	26.2	45.3	19.4	26.8	56.9	22.1	22.9	43.2	9.7	27.8	32.3	24.2	27.3	40.7	28.1
G&M	16.4	34.1	10.5	13.7	32.1	7.7	15.7	43.8	12.3	11.4	30.5	6.7	20.5	24.8	17.9	21.6	32.7	22.3
Happy	22.0	39.4	16.9	19.8	38.4	14.2	21.9	48.7	18.9	18.1	37.0	13.2	24.3	27.9	21.3	24.8	35.6	25.7
PA-CGCD	28.3	46.5	22.7	25.4	44.7	18.4	25.2	55.1	20.9	21.2	41.5	10.2	26.4	31.4	23.7	27.8	40.1	27.2
DEAN	28.9	47.1	23.0	26.3	46.2	18.2	26.1	58.1	19.4	22.1	41.2	12.9	28.1	32.8	28.9	29.1	40.1	30.3
PromptCCD	30.1	48.1	24.5	27.4	46.1	20.3	27.4	57.4	22.1	23.1	44.4	11.4	29.9	34.5	26.4	30.3	42.9	29.9
VB-CGCD	34.2	51.8	26.3	31.7	49.2	23.4	31.6	59.9	26.1	26.3	47.9	15.1	33.2	37.3	29.7	32.3	44.5	31.6
PRISM	**49.3**	**64.9**	**44.2**	**44.0**	**60.9**	**37.0**	**36.9**	**60.0**	**29.1**	**33.3**	**56.5**	**23.5**	**40.1**	**48.9**	**40.1**	**36.4**	**46.1**	**34.1**

Table 3: Component-wise ablation on **Real** → **Painting**.

Components			Real			Painting		
HCS	SAM	IKT	All	Old	New	All	Old	New
✗	✗	✗	54.6	68.7	46.5	28.7	28.1	27.9
✓	✓	✗	58.1	72.9	49.9	35.0	35.9	32.5
✓	✗	✓	56.9	70.2	52.7	33.2	31.8	35.2
✓	✓	✓	**60.9**	**74.1**	**55.1**	**39.2**	**39.0**	**38.2**

Table 4: Comparison of separation strategies on **Real** → **Painting**.

Methods	Real			Painting		
	All	Old	New	All	Old	New
origin image	55.0	68.7	47.2	29.6	28.9	28.3
entropy-based	54.4	69.0	46.7	29.9	29.1	28.6
energy-based	55.8	69.9	48.1	30.6	29.5	29.9
PRISM	**60.9**	**74.1**	**55.1**	**39.2**	**39.0**	**38.2**

HuangmeiSinger: A Dataset and A Branchformer-Diffusion Model for Huangmei Opera Synthesis

Yufeng Qiu
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
2022111052@mail.hfut.edu.cn

Guofu Zhang*
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
zgf@hfut.edu.cn

Zhaopin Su
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
szp@hfut.edu.cn

Yang Zhou
School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China
2023170616@mail.hfut.edu.cn

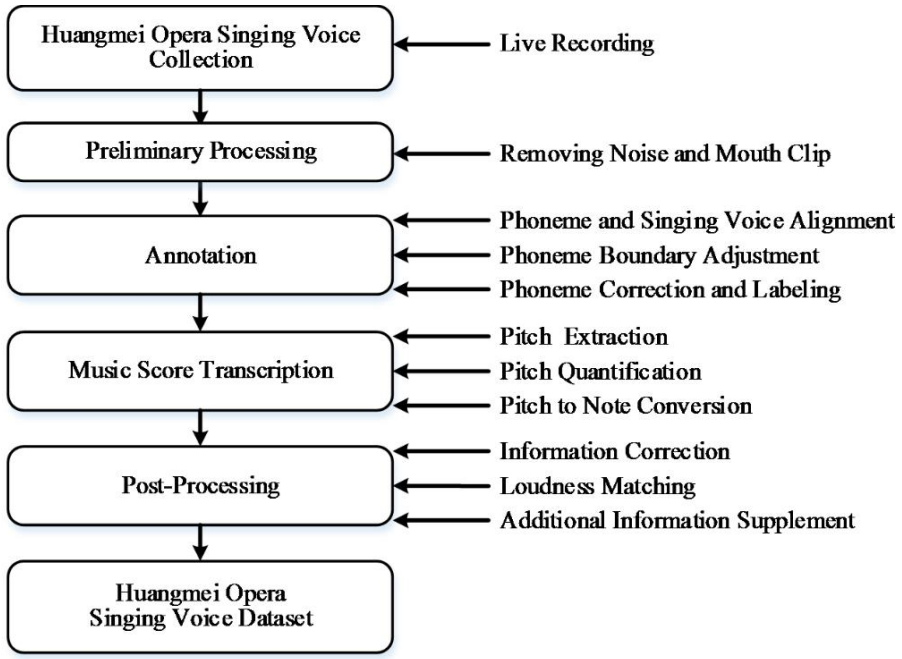
Xiaoyi Bian
College of Humanities and Social Sciences
Anhui Agricultural University
Hefei, Anhui, China
bianxiaoyi@ahau.edu.cn

CAICE 2025

Methods

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

➢ 预处理



➢ 数据集组成（音调、音素、持续时间）

Gender	SingerID	Pitch Range	Hours
Female	Singer1	54-78(D#3,185.00Hz - F#5, 739.99Hz)	1.00
	Singer2	53-76(F3,174.61Hz - E5, 659.25Hz)	0.36
	Singer3	53-77(F3,174.61Hz - F5, 698.46Hz)	0.24
	Singer4	49-73(C#3,138.59Hz – C#5, 554.37Hz)	0.62
Male	Singer5	49-70(C#3,138.59Hz – A#4, 466.16Hz)	0.24
	Singer6	43-69(G2,98.00Hz – A4, 440.00Hz)	1.26

➢ **对齐步骤（音素处理）**

1、歌词转换为音节。根据声母和韵母的规则将获得的音节拆分为音素。
2、通过MFA(Montreal forced aligner model，蒙特利尔强制对准器模型)将音素和录音对齐，结果存在TextGrid。
3、使用Praat注释软件手动校准音素与音节边界。
4、使用Parselmouth获取音调，对音调做后处理。
5、拖音处理

➢ **转录步骤（音调处理）**

1、音调提取
2、量化音调，可以得到音符与MIDI文件的对应编号。$Pt = \frac{\sum_{i=1}^{N} F0_i}{T}$
3、动态拓展（前后各加半拍）

Methods

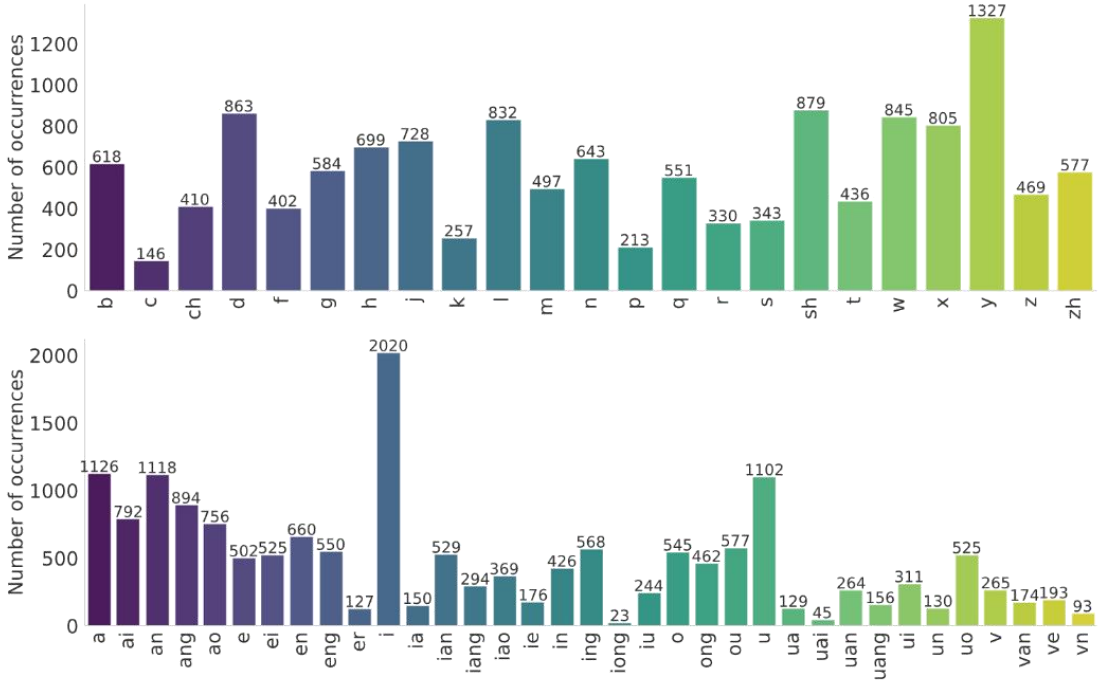


Figure 3: The distribution of phonemes, divided into Shengmu (top) and Yunmu (bottom).

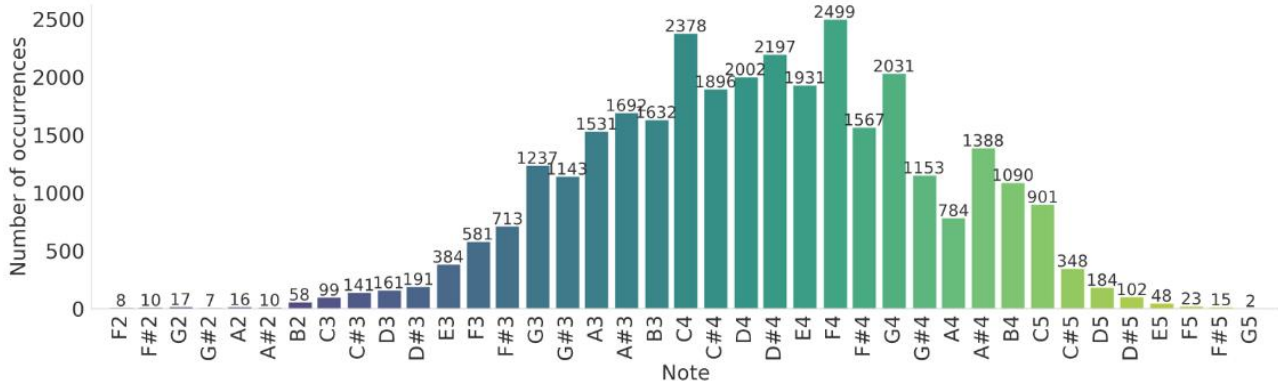
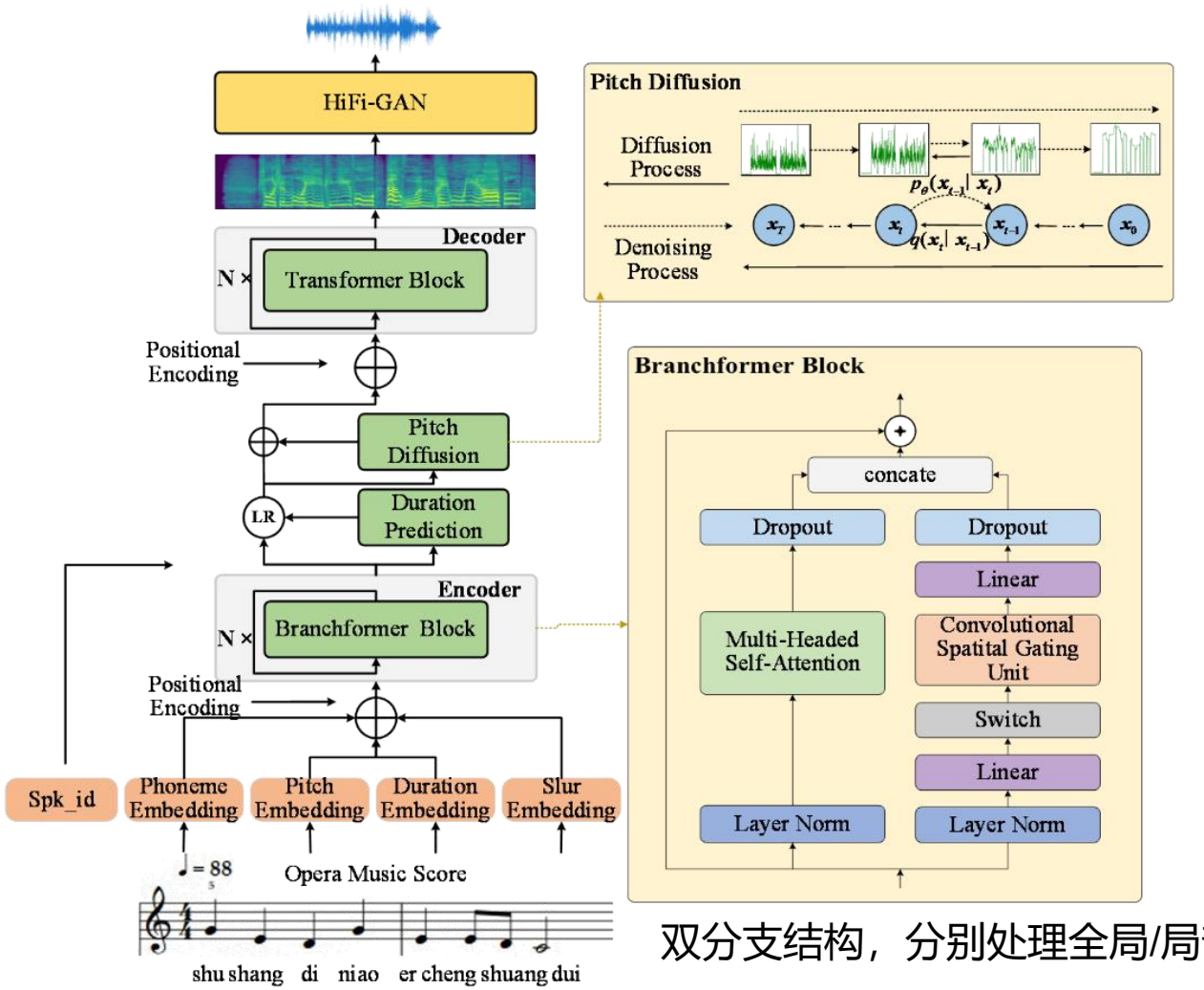


Figure 4: The statistical distribution of pitch in our dataset. Pitch is presented as note.

Methods



双分支结构，分别处理全局/局部信息

- pitch层面

$$\nabla_\theta \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t\right)^2$$

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left[x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, c, t\right)\right] + \sigma_t z$$

用Encoder输出做条件，让音高更贴合文本

- duration层面

$$L_{duration} = \lambda_1 L2_{phoneme} + \lambda_2 L2_{word} + \lambda_3 L2_{sentence}$$

- 梅尔谱层面

$$L_{diff}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\, \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, c, t\right)^2$$

$$L_{total} = L_{duration} + L_{diff} + L_{mel}$$

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 2: Comparative experimental results obtained by the four models in terms of the used evaluation metrics.

Model	MCD↓	F0RMSE↓	DurAcc↑	F0Pcc↑	MOS↑
GT	–	–	–	–	4.36±0.05
FastSpeech2	7.41	24.76	0.838	0.9778	3.54±0.04
DiffSinger	**7.15**	24.51	0.832	0.9784	3.67±0.05
VISinger2	7.69	25.57	0.890	0.9133	3.86±0.04
Ours	7.28	**21.23**	**0.895**	**0.9832**	**3.93±0.05**

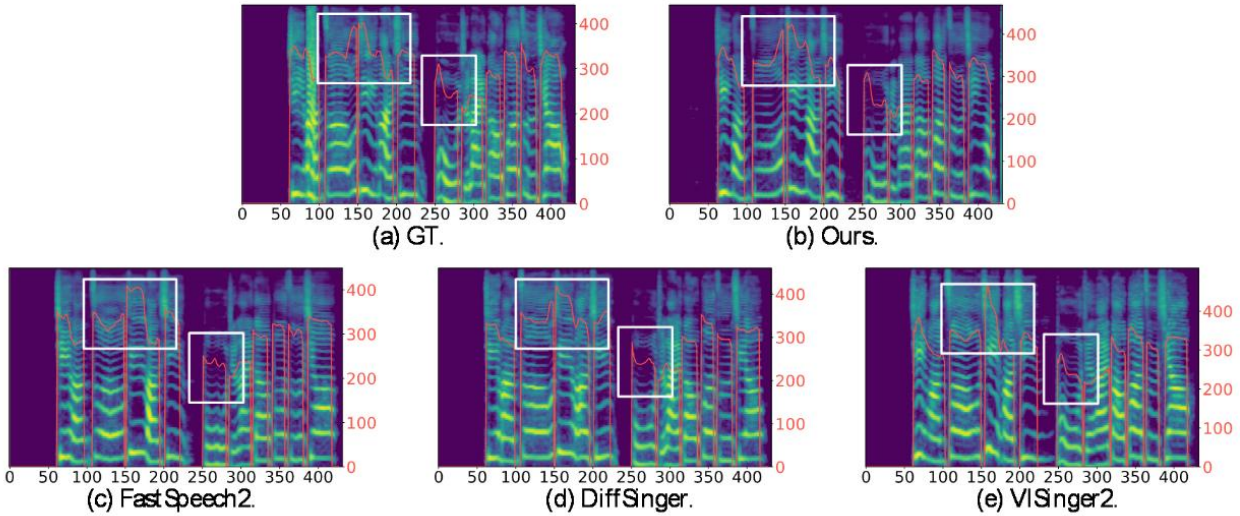


Figure 6: Visualization of the pitch contour and mel-spectrogram of GT and different models (one case study).

- **MCD（Mel-Cepstral Distortion，梅尔倒谱失真）**：衡量生成音频的音色和真实音频的音色有多接近
- **F0 RMSE（基频均方根误差）**：对比生成音频与真实音频的 F0（音高） 差异
- **DurAcc（Duration Accuracy，时长准确率）**：测量生成的音素/音节时长与真实唱法时长的吻合程度
- **F0 PCC（F0 Pearson Correlation Coefficient，基频皮尔逊相关系数）**：衡量生成的 F0 曲线和真实的 F0 曲线形状是否一致

Idea

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

数据：歌词、音频（清唱）

预处理（对齐）：
1、文字转拼音
2、通过MFA对齐拼音和音频
3、通过Praat精细标注
4、使用Parselmouth获取音调F0,并对其优化（可选）

戏曲合成步骤
1、声学建模（**核心**）
2、声码器生成（目前主流：HiFi-GAN/BigVGAN）

可做的方向：
1、纯戏曲数据+戏曲模型（数据体量：4-5h）
2、方言数据+戏曲数据+戏曲模型，从带有方言的戏曲角度切入

Thanks