Advanced Voice Cloning

2022-11-13

候选方案汇总

- Zero-Shot 音色克隆
  - Global and Local Factors: RetrieverTTS
  - Fine-Grained Speaker Embedding

借鉴网络结构，应用时与Few-Shot 微调模型相结合，从而达到更优的效果

- 基于 Adapter 的音色克隆
  - 基于 Adapter 的 TTS 工作
  - 思考：Conditional Adapter

- SOTA: Scaling & Shifting Features

Zero-Shot 音色克隆: RetrieverTTS

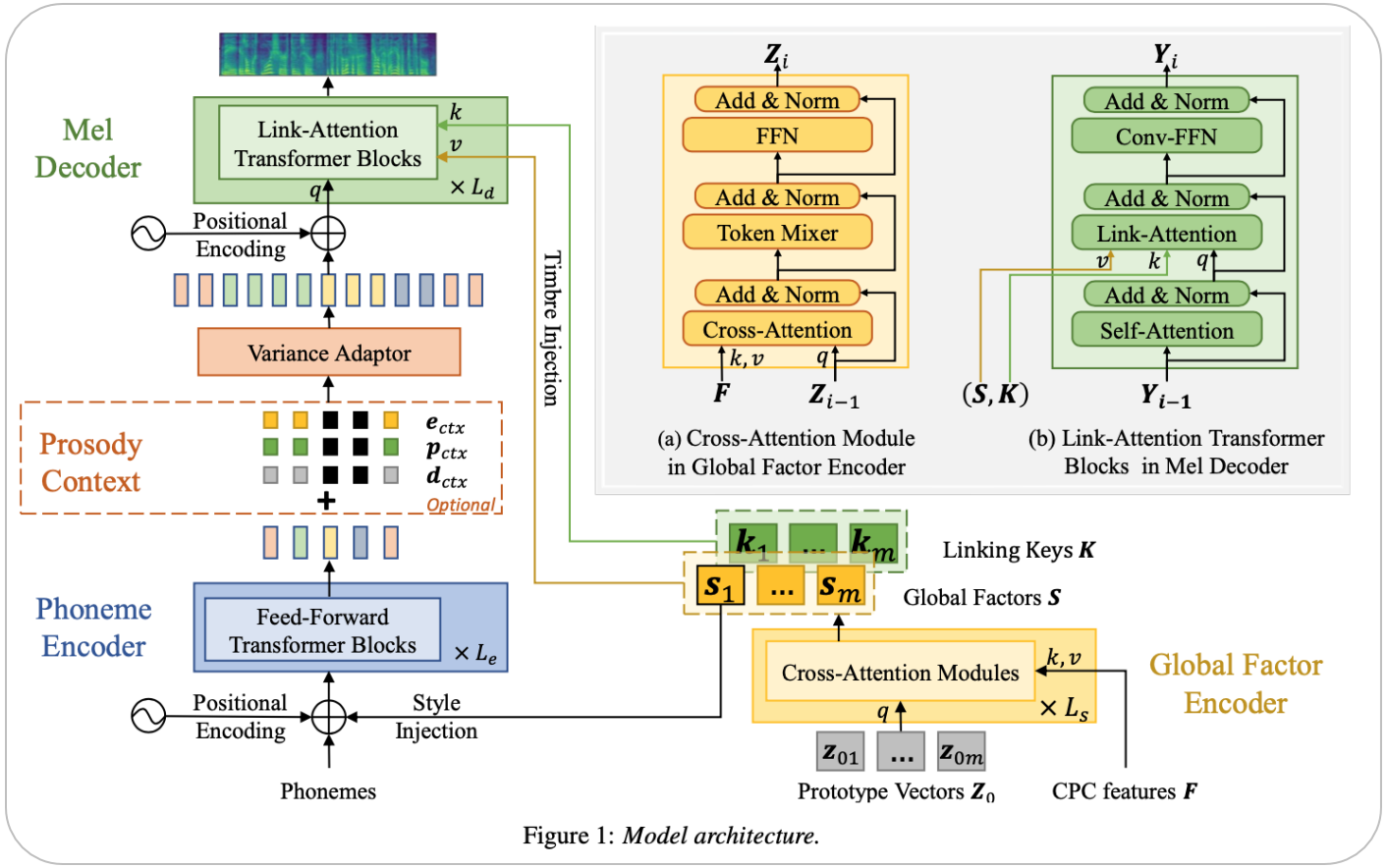


Figure 1: *Model architecture.*

requires whole sentence generation. We also find that our system's SMOS score largely outperforms the state-of-the-art zero-shot voice adaptive TTS method, which leverages an explicit speaker-related meta-learning objective. This indicates that our system is powerful enough to handle the personalized TTS task, even though our design does not fully aim at this goal.

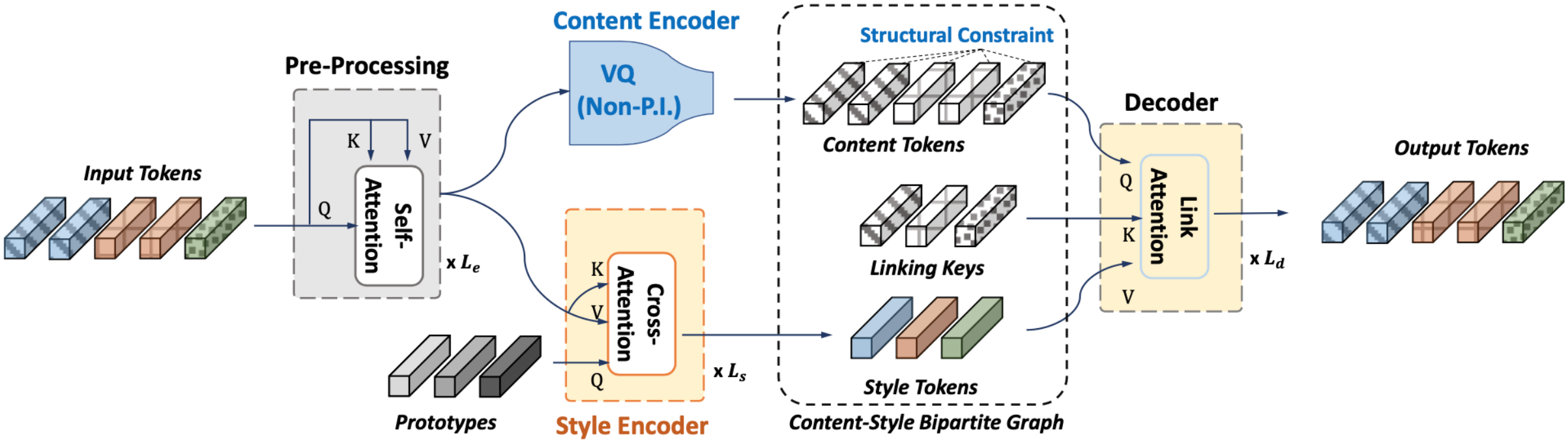
	MOS@long	SMOS
Ground-Truth	4.20 ± 0.14	4.45 ± 0.19
C.Tang et al. [4]	2.73 ± 0.25	—
Meta-StyleSpeech [12]	3.88 ± 0.20	2.91 ± 0.31
RetrieverTTS (Ours)	$\mathbf{3.95} \pm \mathbf{0.17}$	$\mathbf{3.70} \pm \mathbf{0.25}$

RetrieverTTS 用于音色克隆

- 复现模块
  - **CPC 特征**
  - **Cross Attention**
  - **Token Mixer ?**
  - **Link Attention**
- 去除模块 (暂时不实现语音编辑)
  - Prosody Context

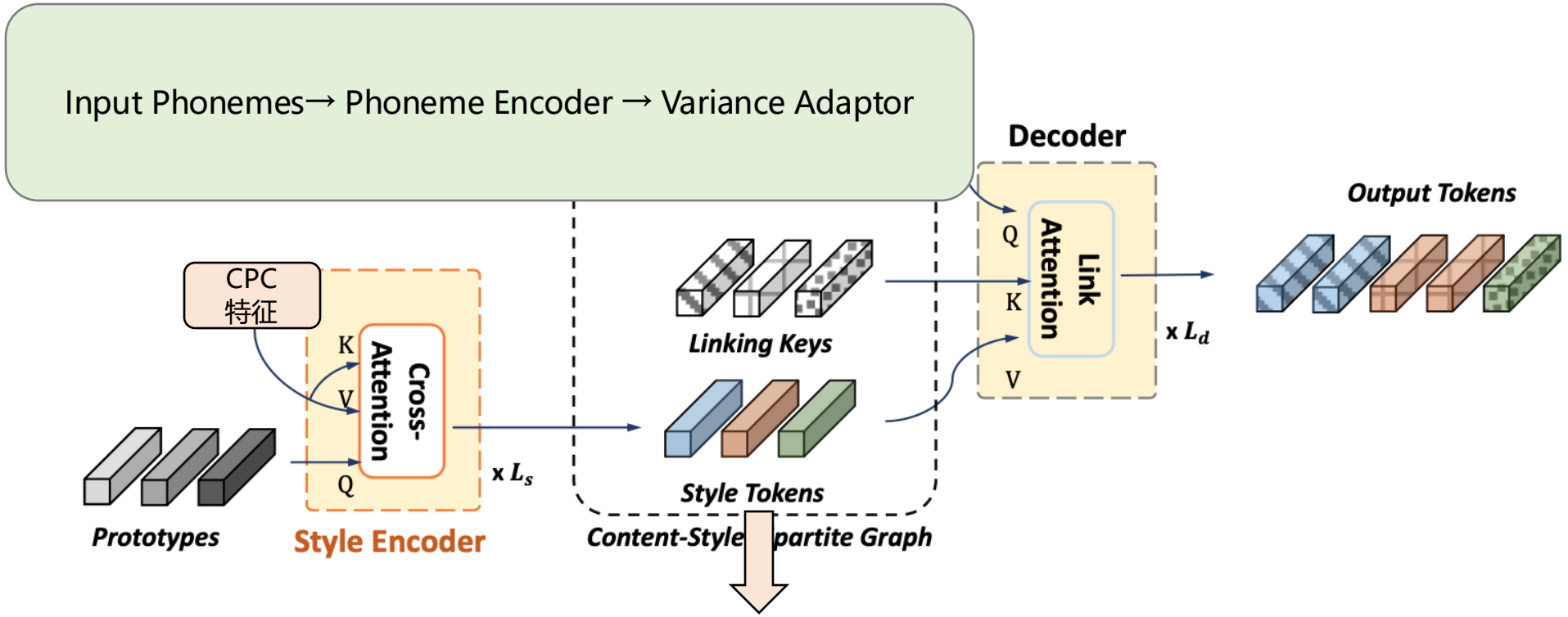
- ICLR 2022: Retriever: Learning Content-Style Representation as a Token-Level Bipartite Graph

Zero-Shot 音色克隆: RetrieverTTS 与 Retriever



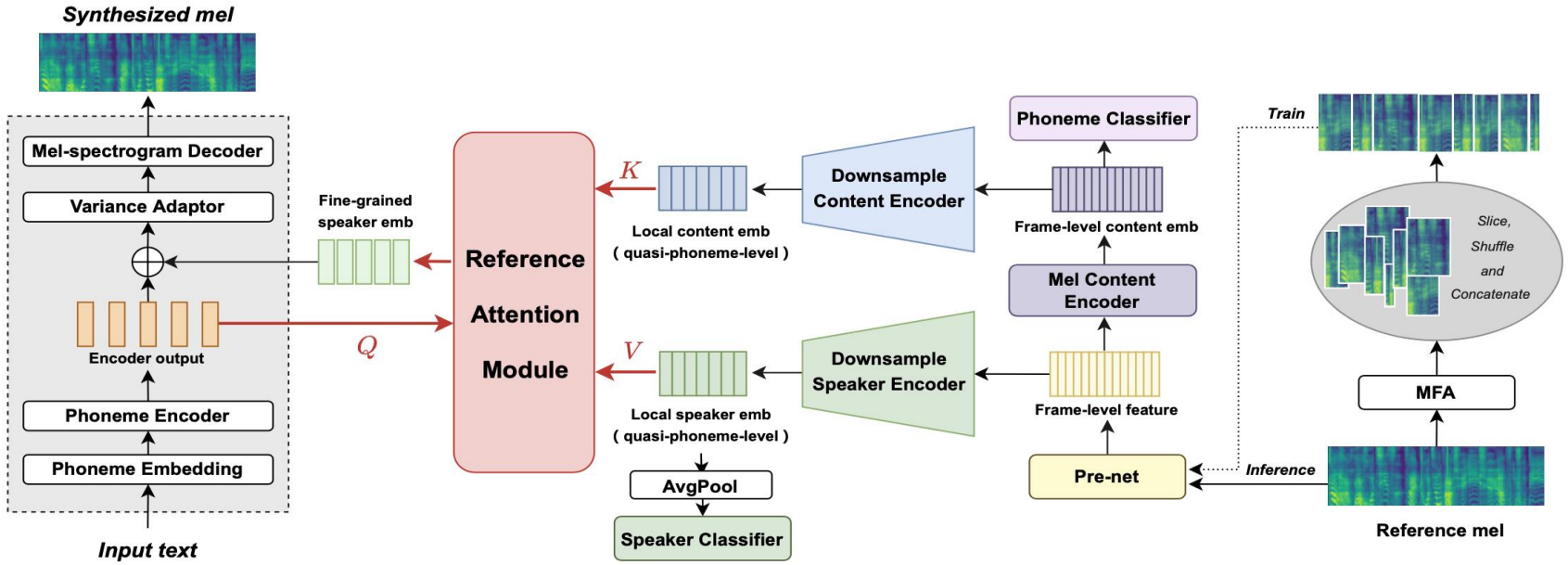
- ICLR 2022: Retriever: Learning Content-Style Representation as a Token-Level Bipartite Graph

Zero-Shot 音色克隆: RetrieverTTS 与 Retriever

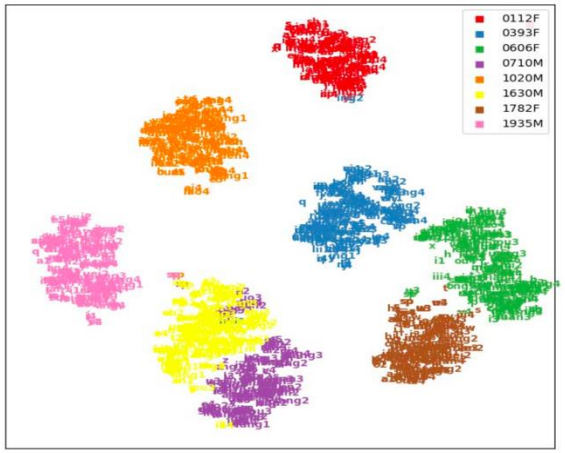


Zero-Shot 音色克隆

Content-Dependent Fine-Grained Speaker Embedding (CDFSE)



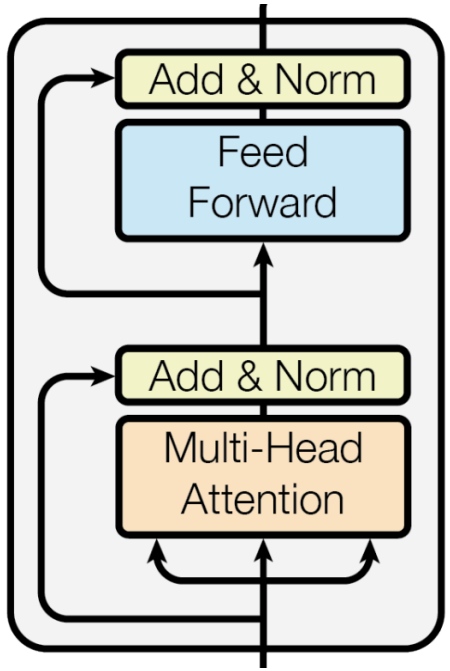
Metric	Model	seen speakers	unseen speakers
MOS	GSE	3.50 ± 0.16	3.56 ± 0.12
	CLS	3.51 ± 0.14	3.53 ± 0.11
	Attentron*	3.63 ± 0.16	3.57 ± 0.13
	CDFSE	3.59 ± 0.17	3.54 ± 0.12
SMOS	GSE	3.89 ± 0.14	3.08 ± 0.14
	CLS	3.79 ± 0.16	3.12 ± 0.14
	Attentron*	4.04 ± 0.17	3.29 ± 0.13
	CDFSE	$\mathbf{4.11 \pm 0.15}$	$\mathbf{3.51 \pm 0.14}$

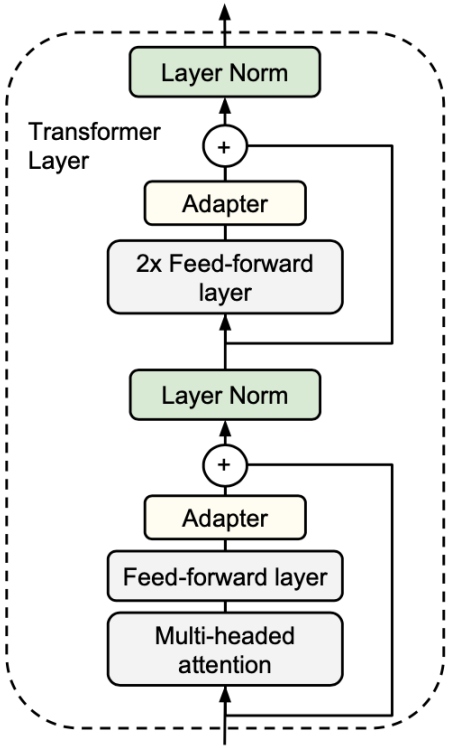


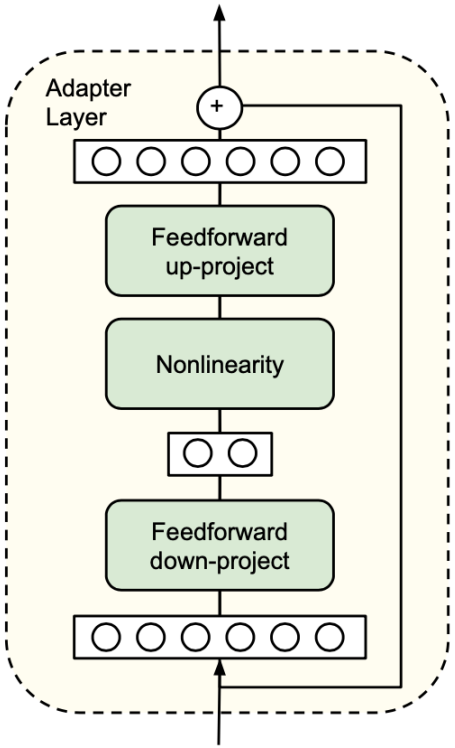
- ICML 2019: Parameter-Efficient Transfer Learning for NLP

基于 Adapter 的方法

Adapter 的提出: Parameter-Efficient Transfer Learning for NLP





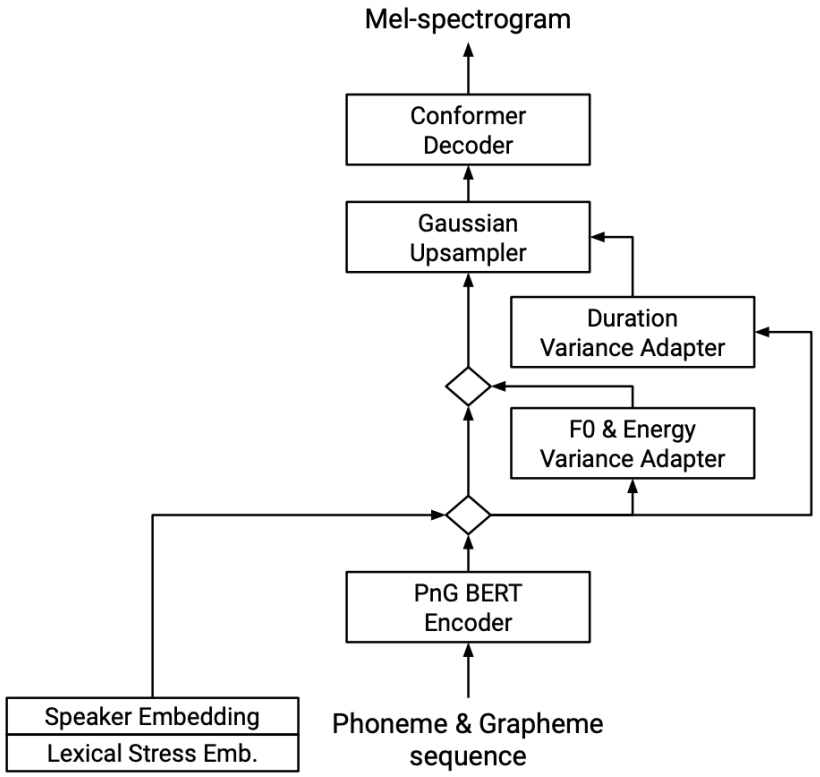


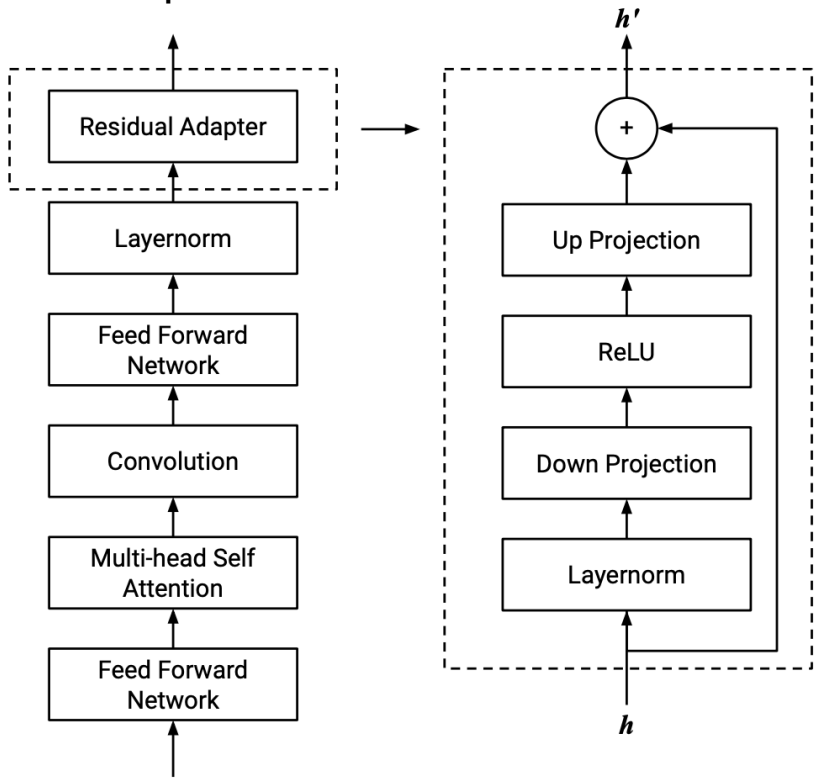
- finetune 时只更新 Adapter + LayerNorm 部分的参数

- ICASSP 2023: Residual Adapters for Few-Shot Text-to-Speech Speaker Adaptation

基于 Adapter 的方法

论文 1: Residual Adapters for Few-Shot Text-to-Speech Speaker Adaptation

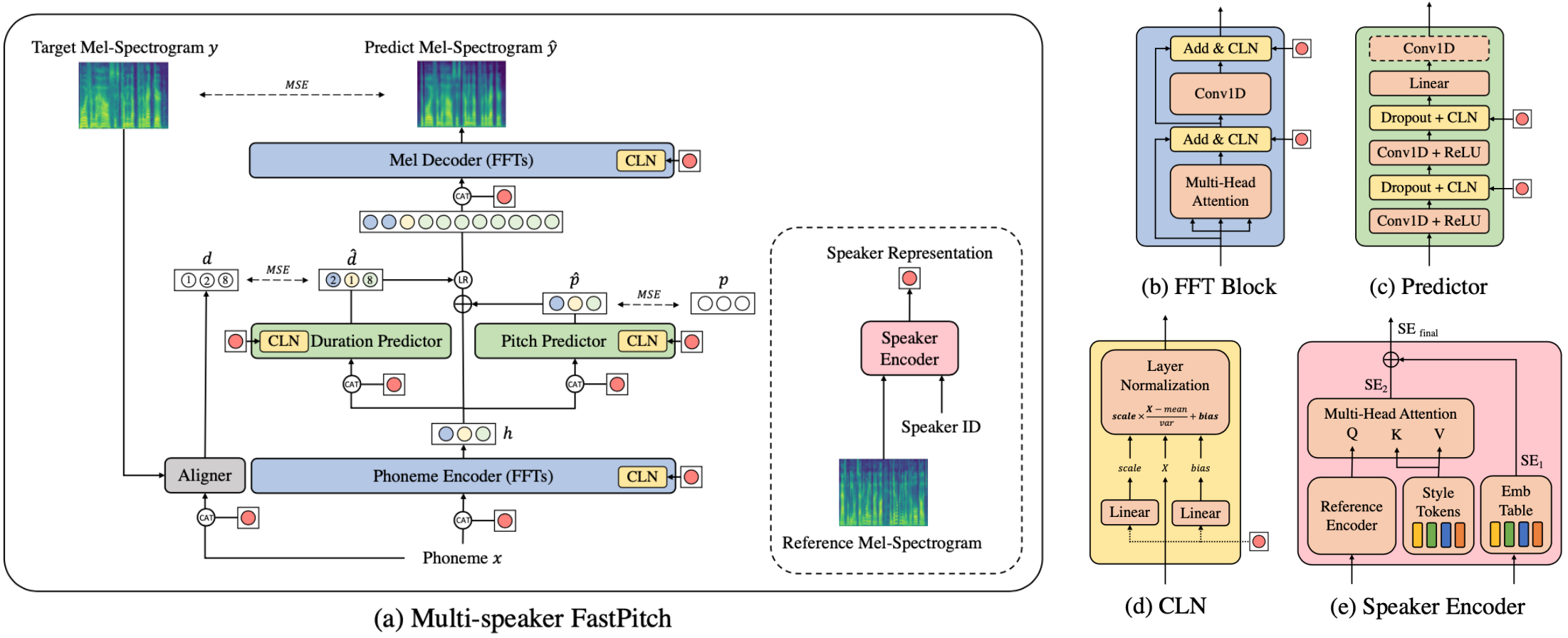




$$h' = h + \mathrm{ReLU}\left(\mathrm{LayerNorm}\left(h\right) W_{\mathrm{down}}\right) W_{\mathrm{up}}$$

基于 Adapter 的方法

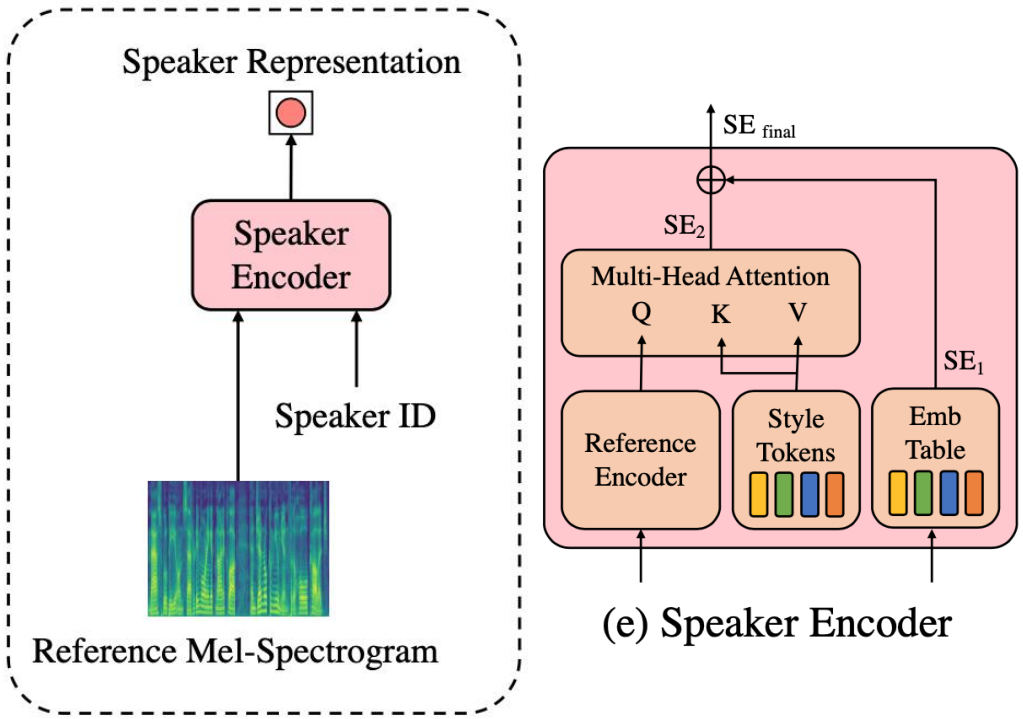
论文 2: Adapter-Based Extension of Multi-Speaker Text-to-Speech Model for New Speakers



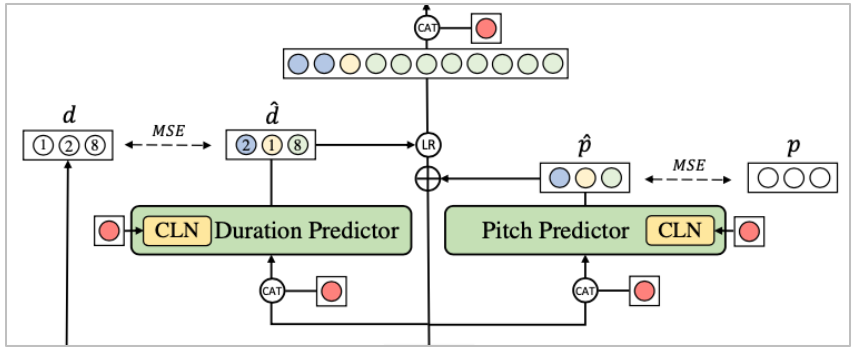
基于 Adapter 的方法

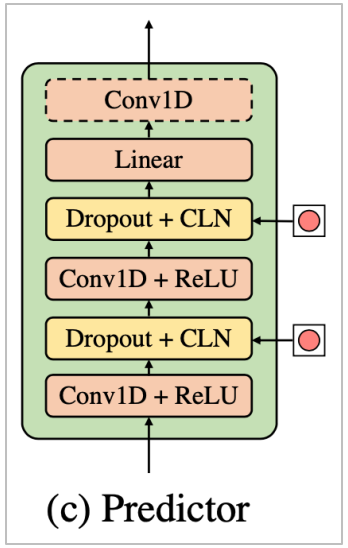
论文 2: Adapter-Based Extension of Multi-Speaker Text-to-Speech Model for New Speakers

借鉴点一



借鉴点二

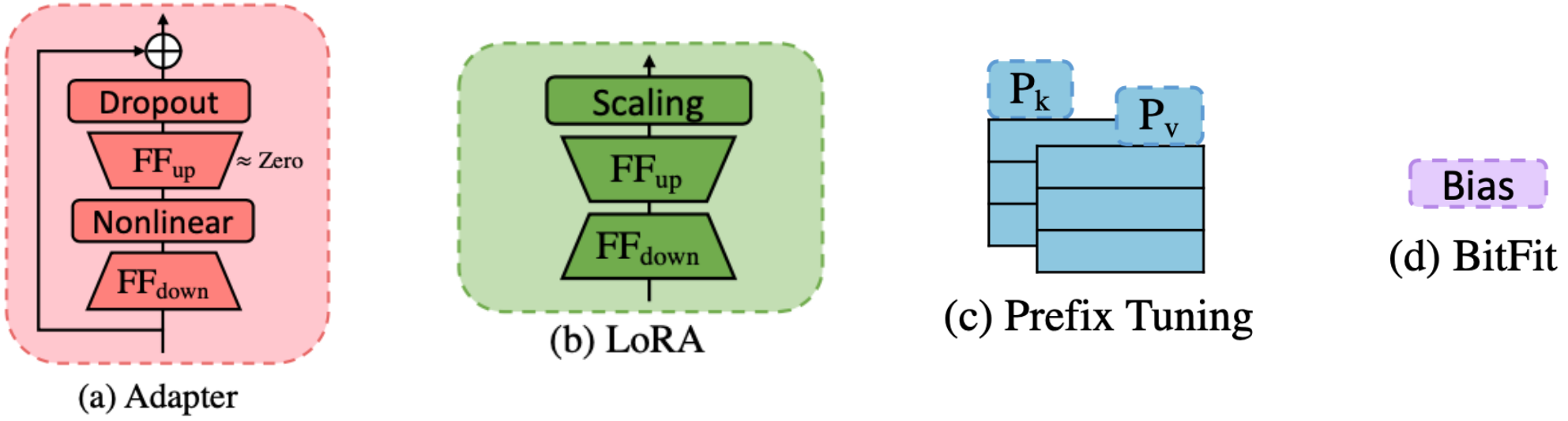




基于 Adapter 的方法

论文 2: Adapter-Based Extension of Multi-Speaker Text-to-Speech Model for New Speakers

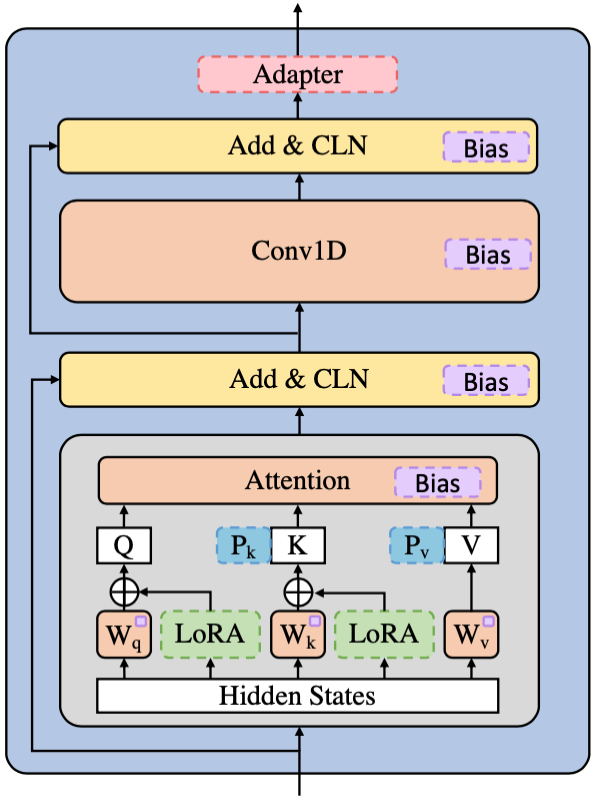
- 统一框架下的 Adapter 及变体



基于 Adapter 的方法

论文 2: Adapter-Based Extension of Multi-Speaker Text-to-Speech Model for New Speakers

- 统一框架下的 Adapter 及变体



Method	SECS ↑	CFSD ↓	MSE_P ↓	MSE_D ↓	Params
Different parameter-efficient methods					
BitFit	0.452	56.4	71.0	**19.1**	2.2M
PrefixTuning (FFTs)	0.067	83.2	75.6	22.7	0.6M
LoRA (FFTs)	0.141	62.7	77.7	22.3	2.8M
Adapter (FFTs)	**0.568**	**30.0**	**65.9**	21.3	2.4M
Different speaker-related modules					
Adapter (FFTs/Predictors/Aligner)	0.575	28.3	**62.7**	16.9	3.5M
+ speaker embedding	**0.586**	**27.2**	63.6	**16.2**	3.5M
+ speaker embedding + CLN	0.540	46.9	66.8	17.2	7.8M
speaker embedding + CLN [13]	0.513	53.5	68.0	21.7	4.3M
Full fine-tuning	0.604	31.0	73.8	19.7	53.4M

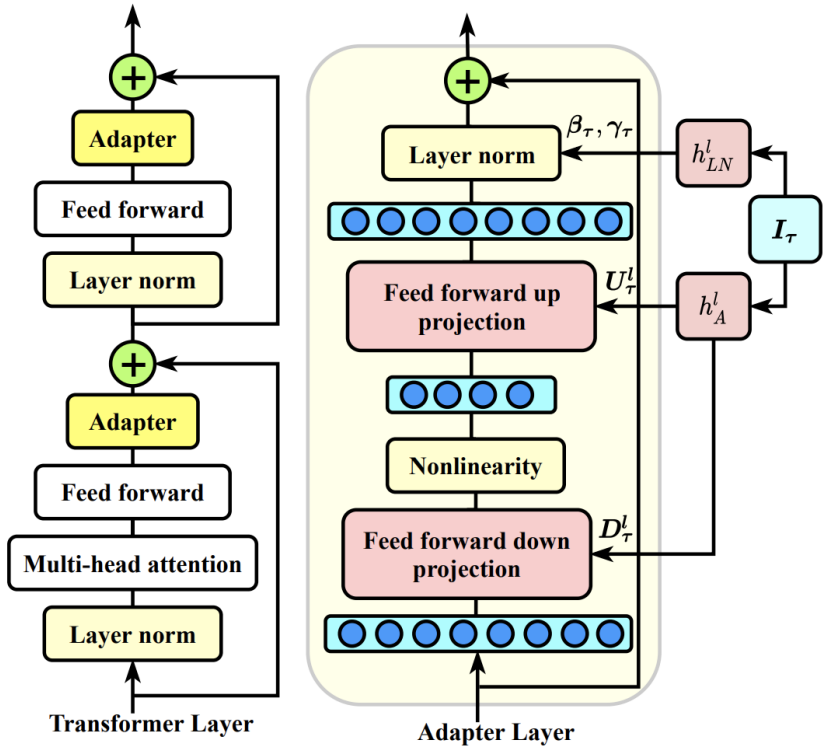
	MOS ↑			SMOS ↑		
Method	**LibriTTS**	**VCTK**	**HiFi-TTS**	**LibriTTS**	**VCTK**	**HiFi-TTS**
Recording	4.16 ± 0.05	4.08 ± 0.05	4.12 ± 0.05	3.90 ± 0.07	3.80 ± 0.07	3.71 ± 0.08
Adapter	**4.10 ± 0.05**	**3.89 ± 0.06**	**3.87 ± 0.06**	**3.44 ± 0.08**	3.28 ± 0.08	3.27 ± 0.08
Full fine-tuning	3.96 ± 0.06	3.88 ± 0.06	3.75 ± 0.07	3.40 ± 0.08	**3.33 ± 0.06**	**3.34 ± 0.08**

- ACL 2021: Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks

基于 Adapter 的方法

Conditional Adapter

参考论文: Parameter-efficient Multi-task Fine-tuning for Transformers via **Shared Hypernetworks**



$$A_\tau^l(\boldsymbol{x}) = LN_\tau^l\left(\boldsymbol{U}_\tau^l(\mathrm{GeLU}(\boldsymbol{D}_\tau^l(\boldsymbol{x})))\right) + \boldsymbol{x}$$

$$(\boldsymbol{U}_\tau^l, \boldsymbol{D}_\tau^l) := h_A^l(\boldsymbol{I}_\tau) = \left(\boldsymbol{W}^{\boldsymbol{U}^l}, \boldsymbol{W}^{\boldsymbol{D}^l}\right)\boldsymbol{I}_\tau$$

$$\boldsymbol{W}^{\boldsymbol{U}^l} \in \mathbb{R}^{(d\times h)\times t}$$

$$\boldsymbol{W}^{\boldsymbol{D}^l} \in \mathbb{R}^{(h\times d)\times t}$$

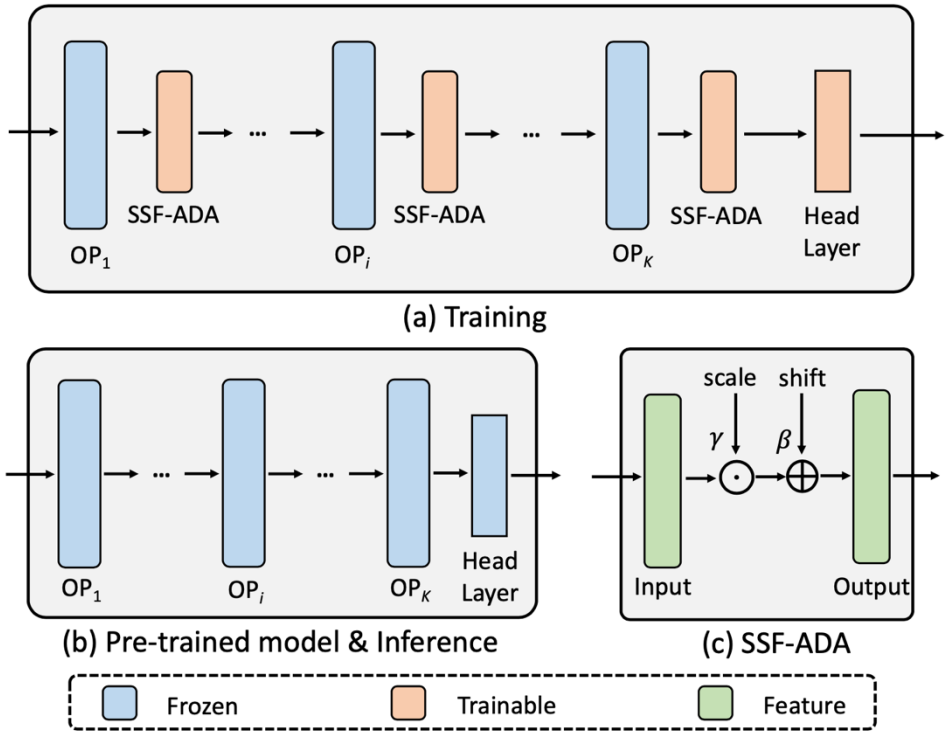
$$(\boldsymbol{\gamma}_\tau^l, \boldsymbol{\beta}_\tau^l) := h_{LN}^l(\boldsymbol{I}_\tau) = \left(\boldsymbol{W}^{\boldsymbol{\gamma}^l}, \boldsymbol{W}^{\boldsymbol{\beta}^l}\right)\boldsymbol{I}_\tau$$

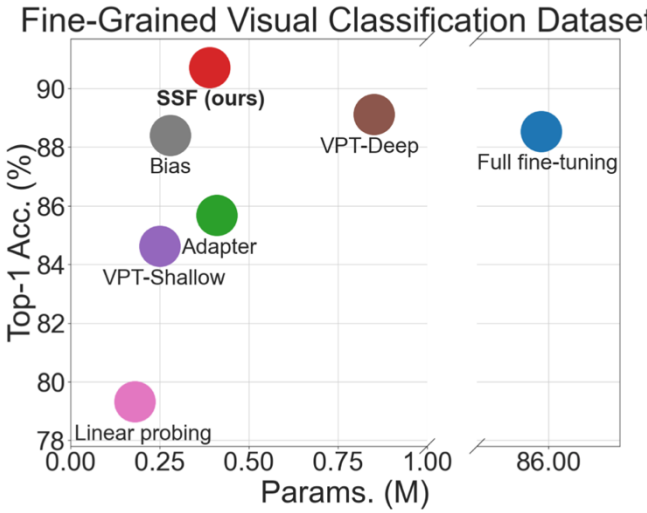
$$\boldsymbol{W}^{\boldsymbol{\gamma}^l} \in \mathbb{R}^{h\times t}$$

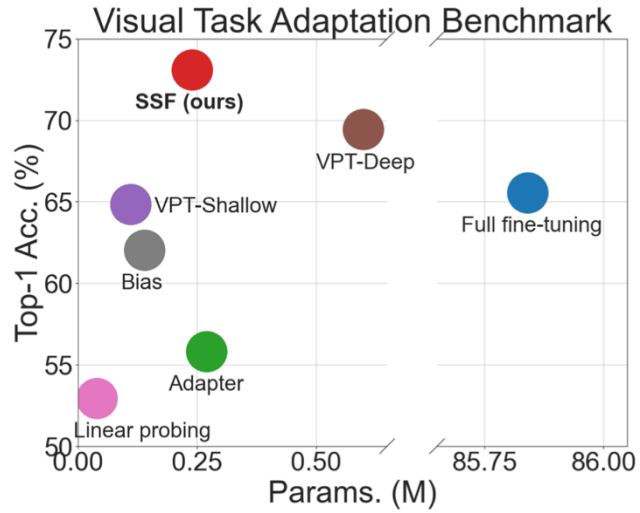
$$\boldsymbol{W}^{\boldsymbol{\beta}^l} \in \mathbb{R}^{h\times t}$$

- NeurIPS 2022: Scaling & Shifting Your Features: A New Baseline for Efficient Model Tuning

SOTA: Scaling & Shifting Features (SSF)







Method \ Dataset	CUB-200 -2011	NABirds	Oxford Flowers	Stanford Dogs	Stanford Cars	Mean	Params. (M)
Full fine-tuning	87.3	82.7	98.8	89.4	84.5	88.54	85.98
Linear probing	85.3	75.9	97.9	86.2	51.3	79.32	0.18
Adapter [36]	87.1	84.3	98.5	89.8	68.6	85.67	0.41
Bias [90]	88.4	84.2	98.8	**91.2**	79.4	88.41	0.28
VPT-Shallow [44]	86.7	78.8	98.4	90.7	68.7	84.62	0.25
VPT-Deep [44]	88.5	84.2	99.0	90.2	83.6	89.11	0.85
SSF **(ours)**	**89.5**	**85.7**	**99.6**	89.6	**89.2**	**90.72**	0.39

方案思考与总结

候选方案

- 按照优先级顺序排列
  - RetrieverTTS
  - (Conditional) Adapter-Based Method
  - Scaling & Shifting Features (SSF)
  - Content-Dependent Fine-Grained Speaker Embedding (CDFSE)