音频事件检测研究汇报

主讲人：张成飞
指导老师：翟克博士

Background

- 音频事件检测是声音模式识别中的一个重要分支
- 能够实现特定场景下的声音事件的监测和分析
- 处理的数据音频不同于语音和音乐

The Purpose

- 并非所有的声音都能通过人耳进行辨别分类
- 处理的数据更多的是噪音
- 机器更加客观，且无间断工作

Datasets Introduction

- ESC-50
  - 包含2000个环境声音集、50[1]个标签类、划分为5个fold
  - 每段音频时长大约5s，采样率为44.1kHz
  - 音频较清晰
- UrbanSound8K
  - 包含8000个环境声音集、10[2]个标签类、划分为10个fold
  - 每段音频时长大约4s，采样率分布在8kHz-96kHz
  - 音频接近真实场景

[1] 狗吠、打喷嚏、鼓掌、咳嗽、抽水马桶、洗衣机、鼠标点击、洗衣机、直升机……;
[2] 空调、汽车鸣笛、儿童嬉戏、狗吠、钻头、发动机、枪炮、电钻、警报、街头音乐;

Between Class Learning - A Novel Data Augmentation Approach

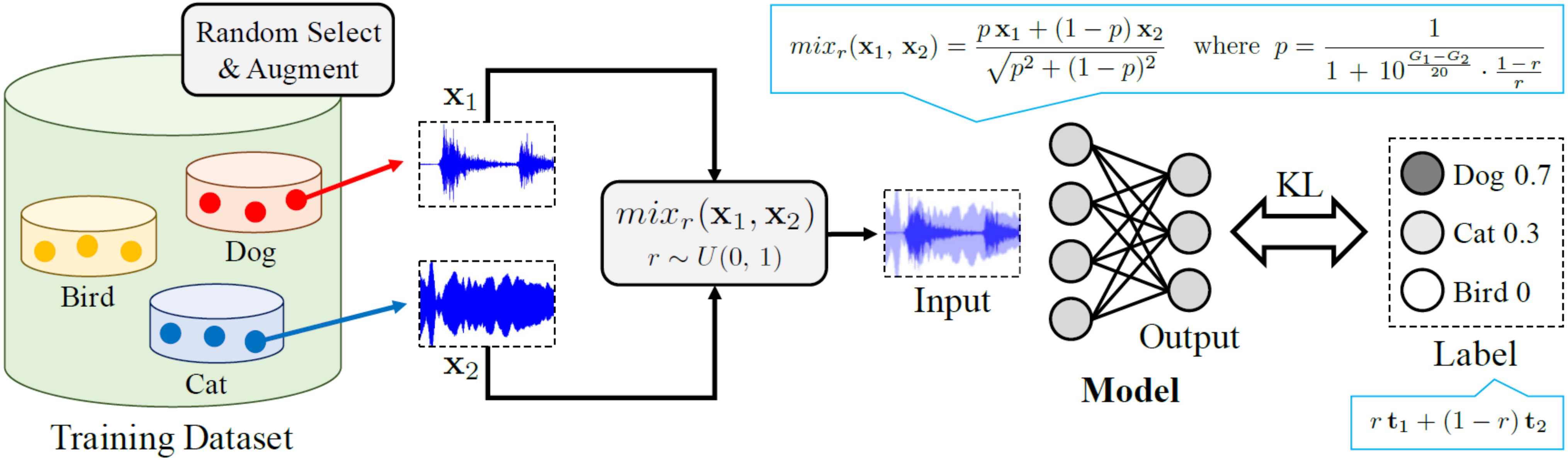
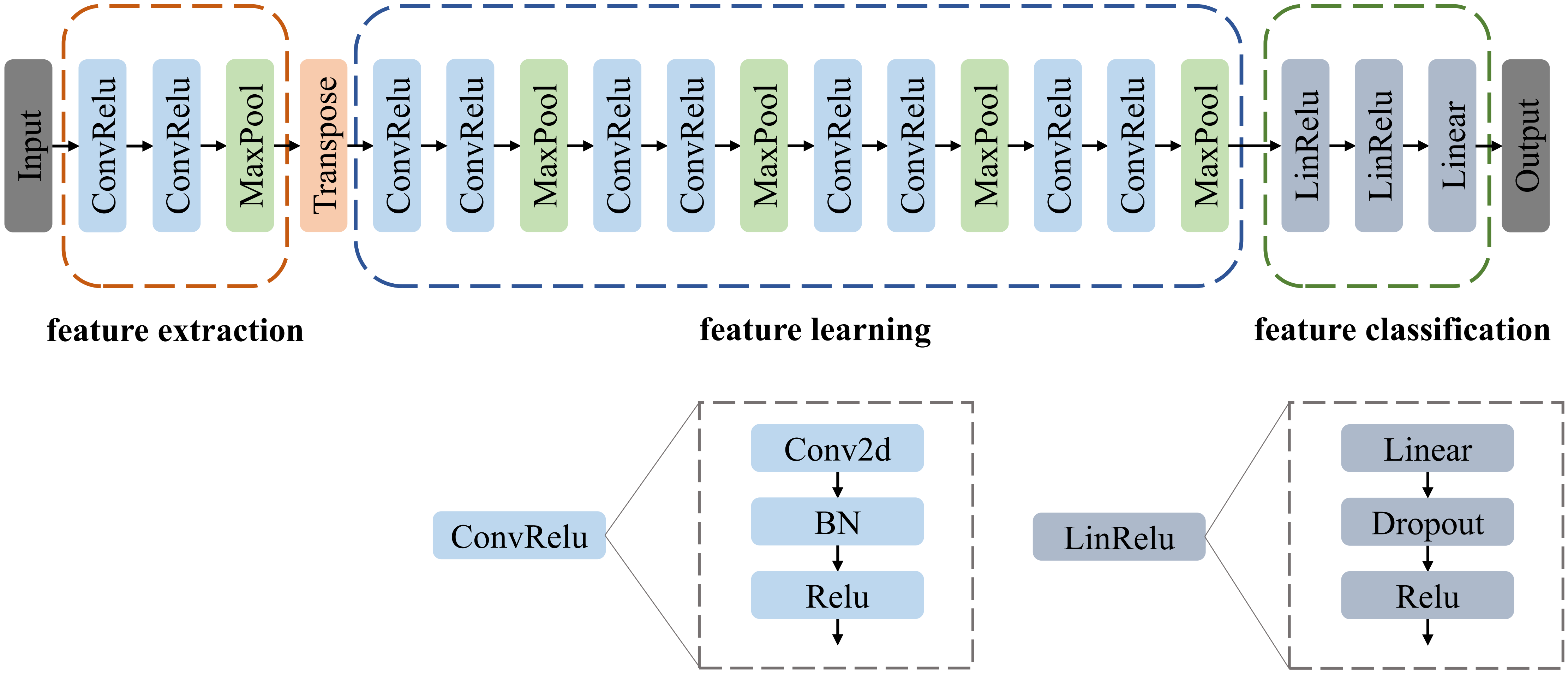


Figure 1: Pipeline of BC learning. We create each training example by mixing two sounds belonging to different classes with a random ratio. We input the mixed sound to the model and train the model to output the mixing ratio using the KL loss.

Baseline Model - EnvNet2

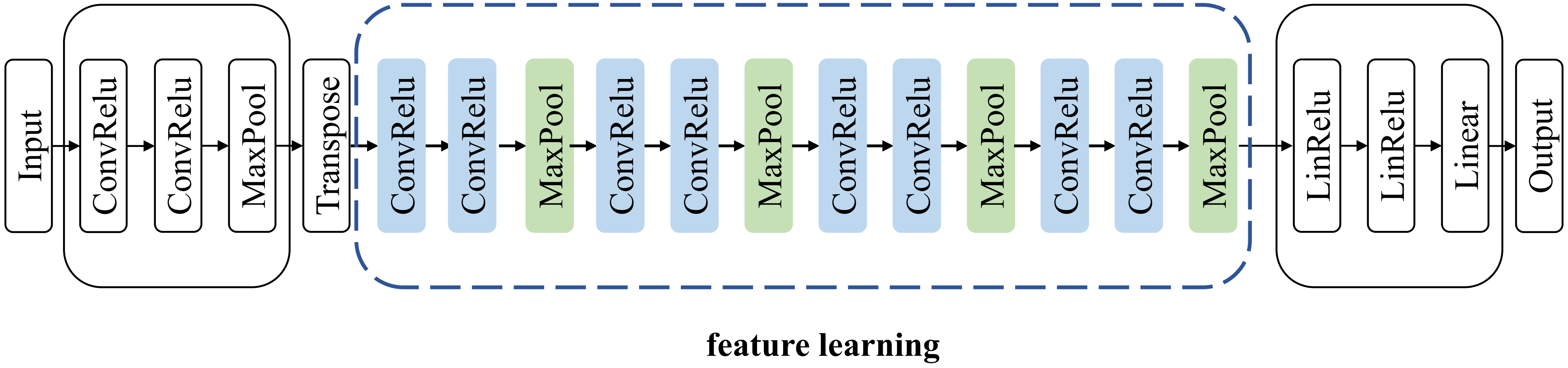


Model Result

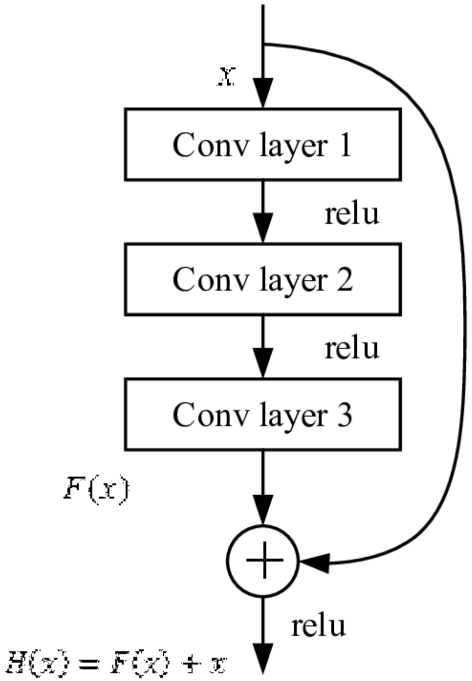
Model Name	Parameters	Input Feature	ESC50	UrbanSound8K
EnvNet2	101.25M	wav	78.50 [2]	76.60
EnvNet2+BCLearning [1]			84.70 (+6.20) [3]	78.30 (+1.70)

[1] 后述EnvNet2均为 EnvNet2+BCLearning，且提出模型均采用BCLearning
[2] 模型在5个fold上预测的最大准确率的平均值
[3] 相对前一模型准确率的提升值

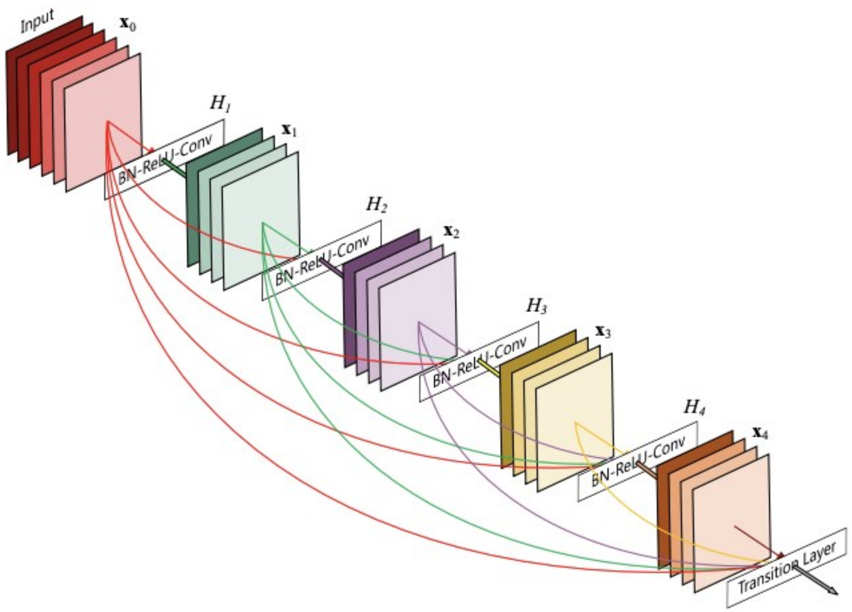
Rethink Baseline Model



ResNet or DenseNet

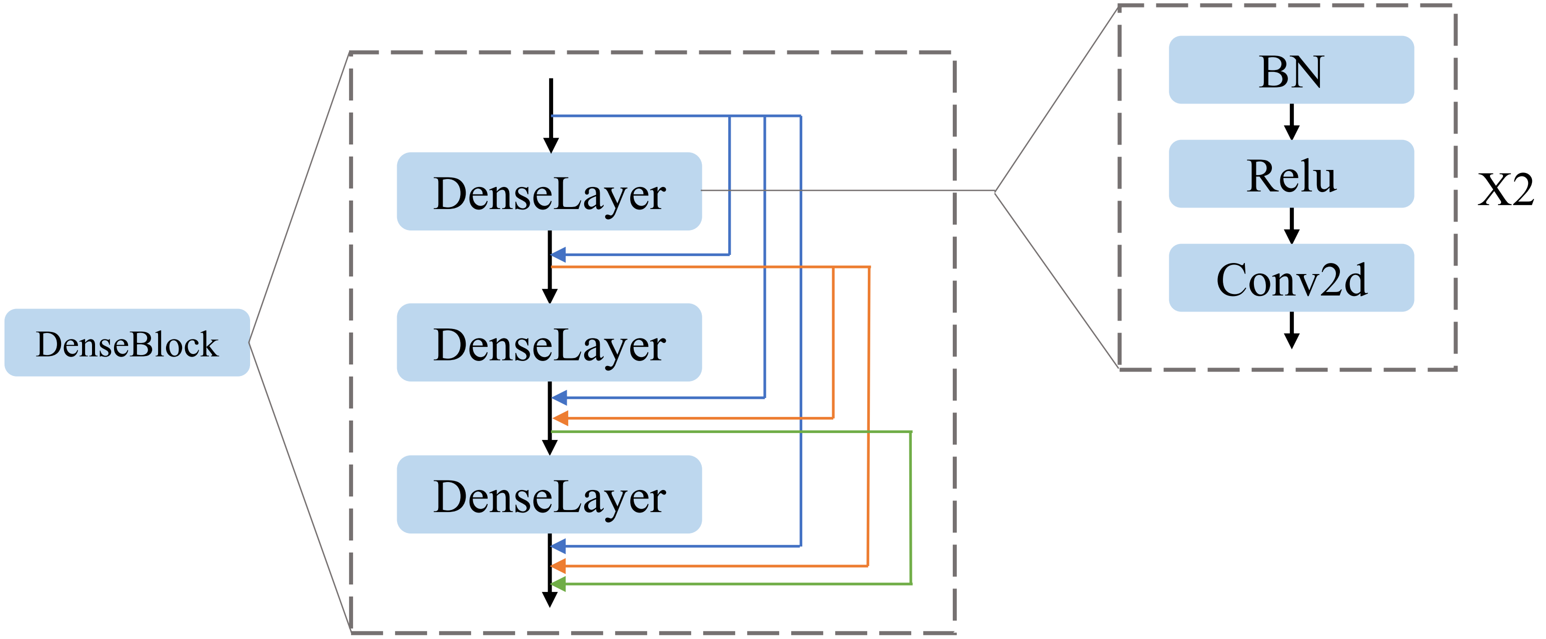


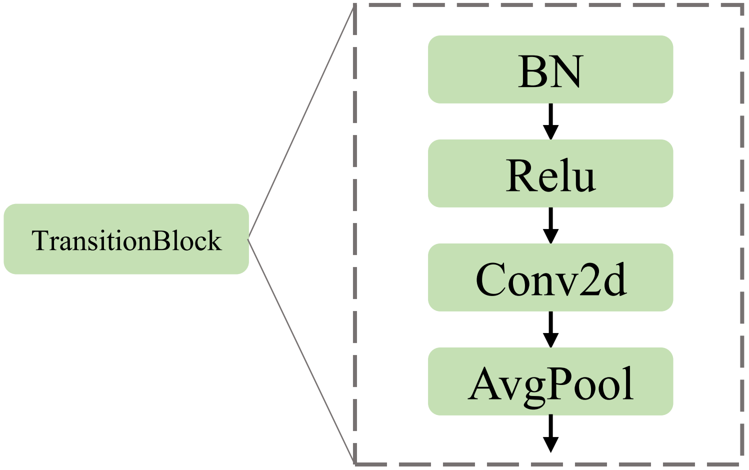
ResNet Module

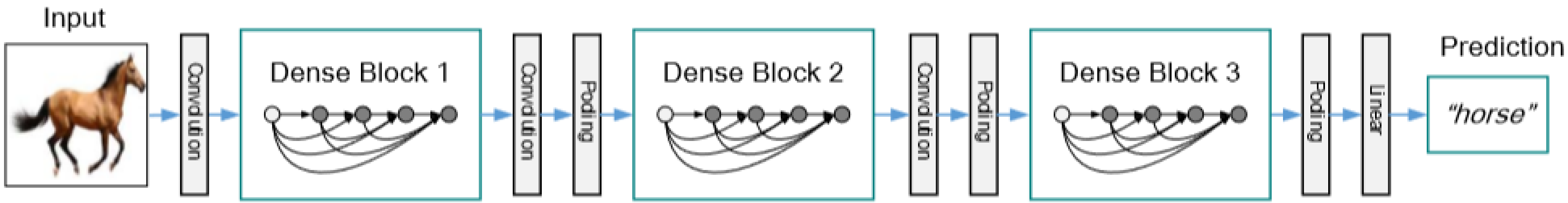


DenseNet Module

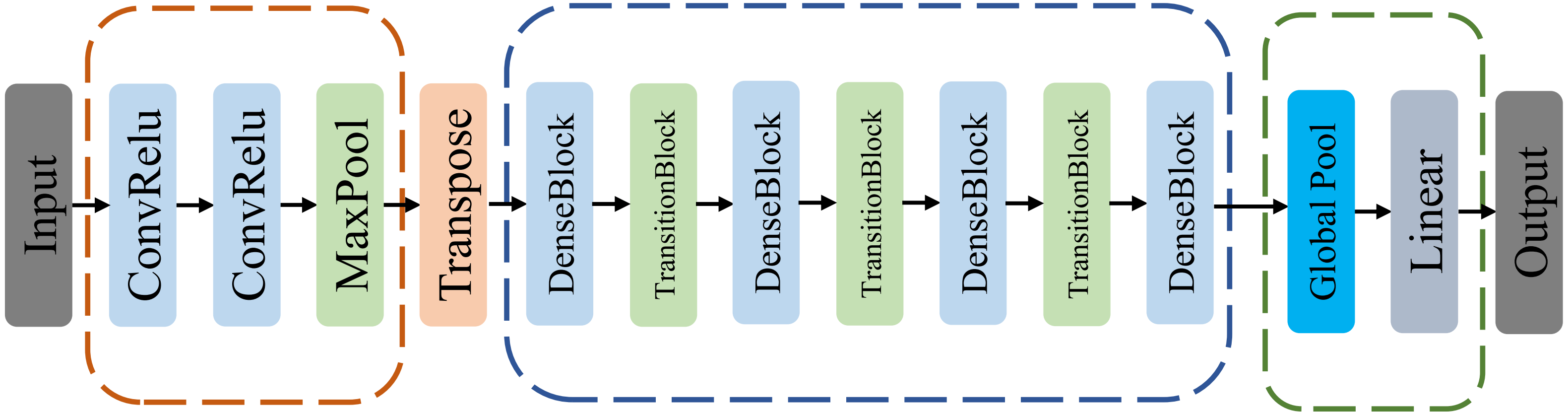
DenseNet Model Details







Proposed1 - use DenseNet



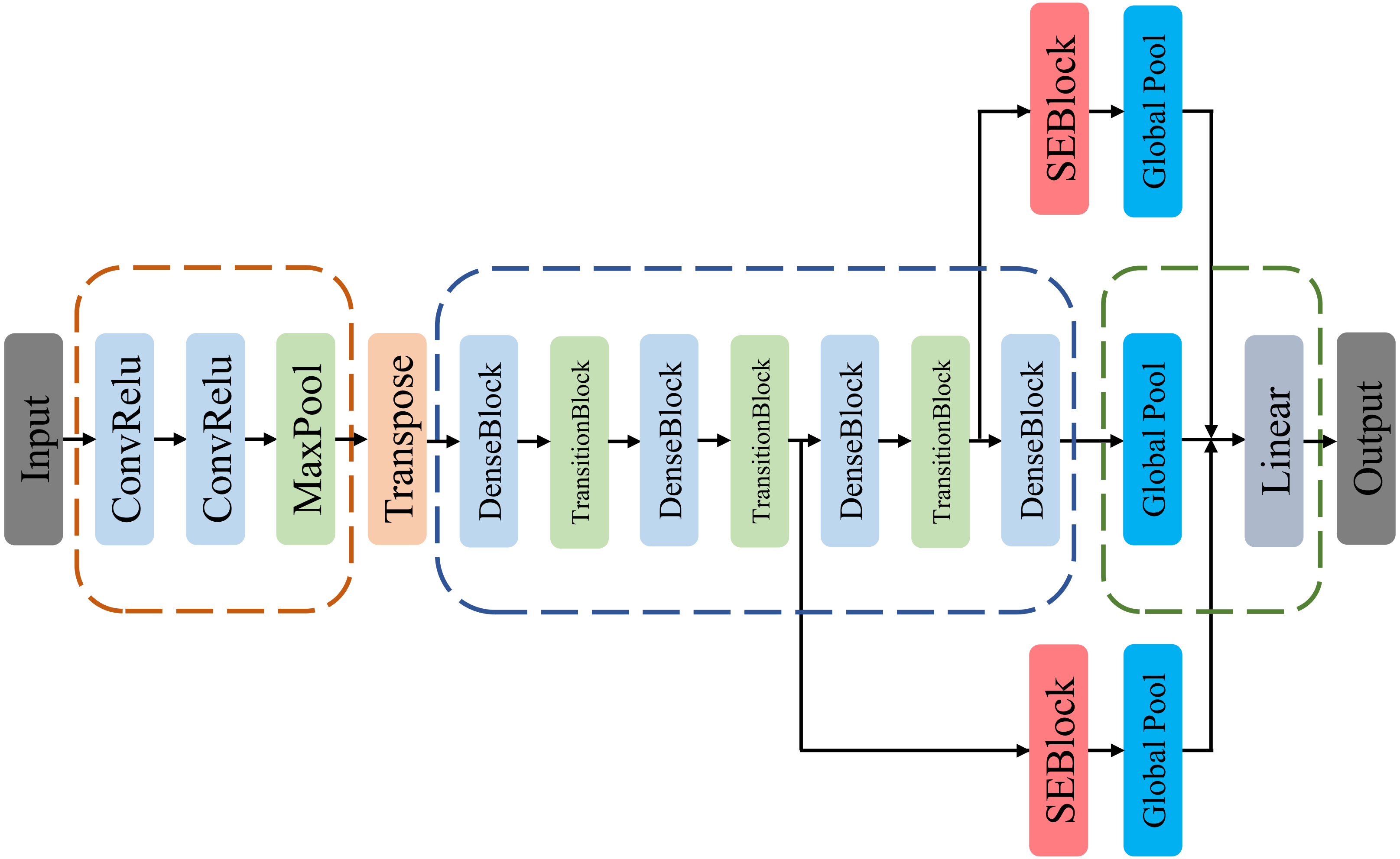
feature learning

Model Result

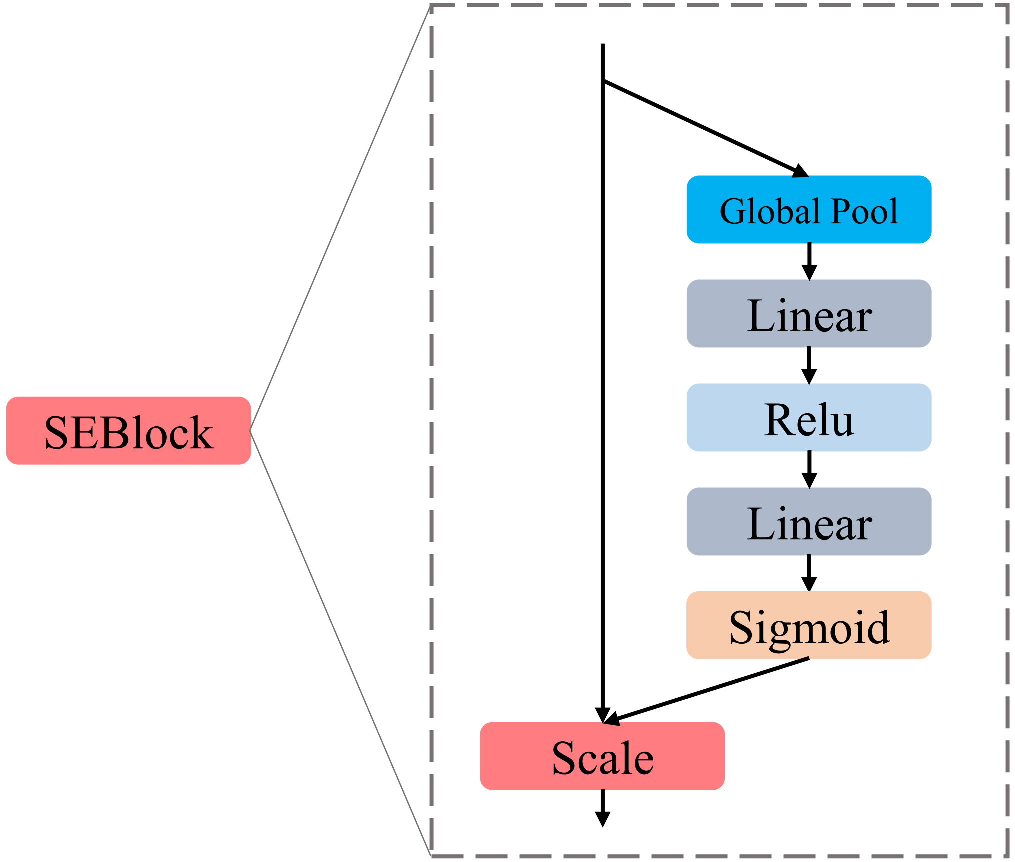
Model Name	Parameters	Input Feature	ESC50
EnvNet2	101.25M	wav	84.70
Proposed1	7.03M (-93.06%) [1]		87.85 (+3.15)

[1] 相对前一模型参数下降百分比

Proposed2 - use Multi-Scale and SEBlock



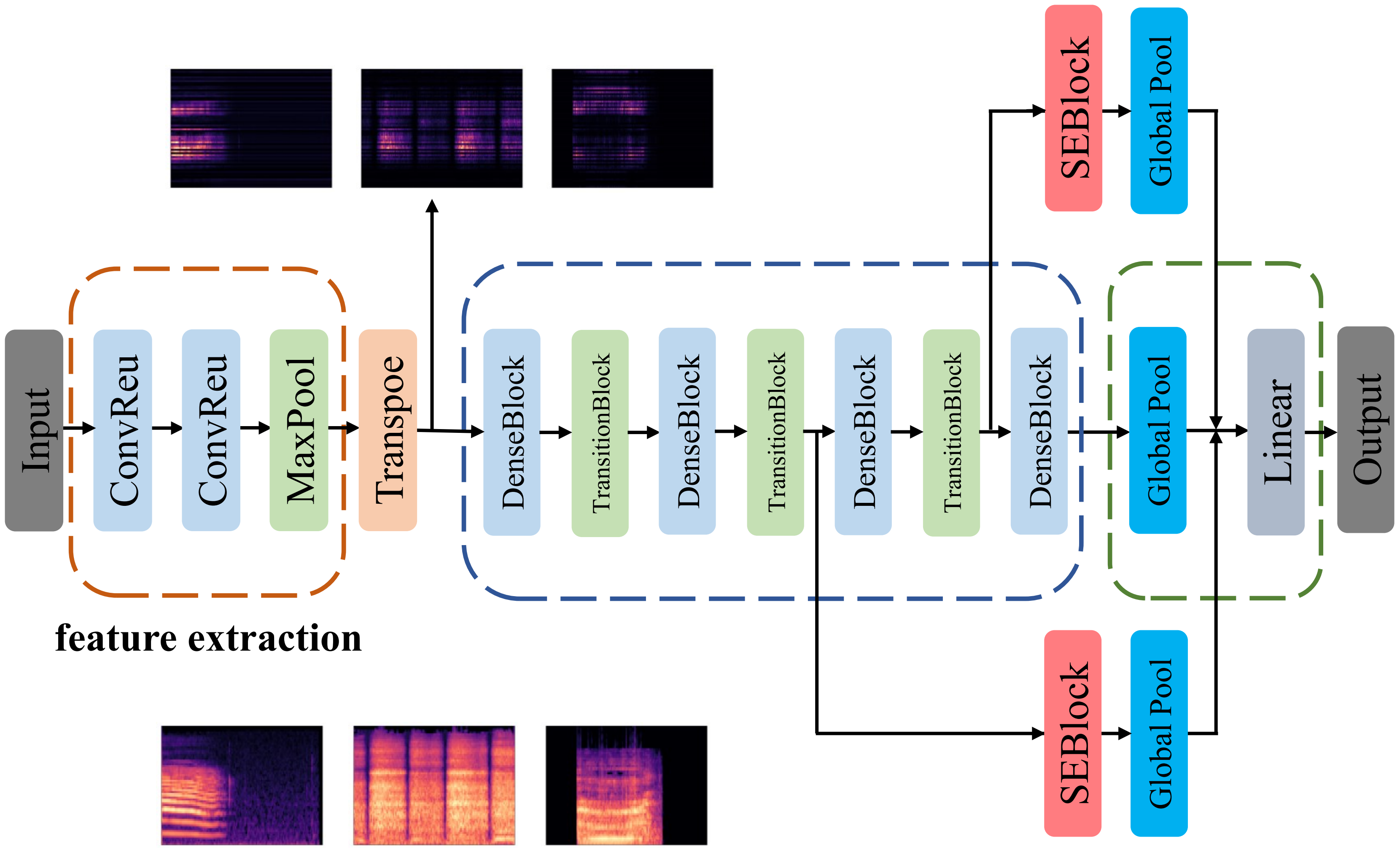
SEBlock Model Details



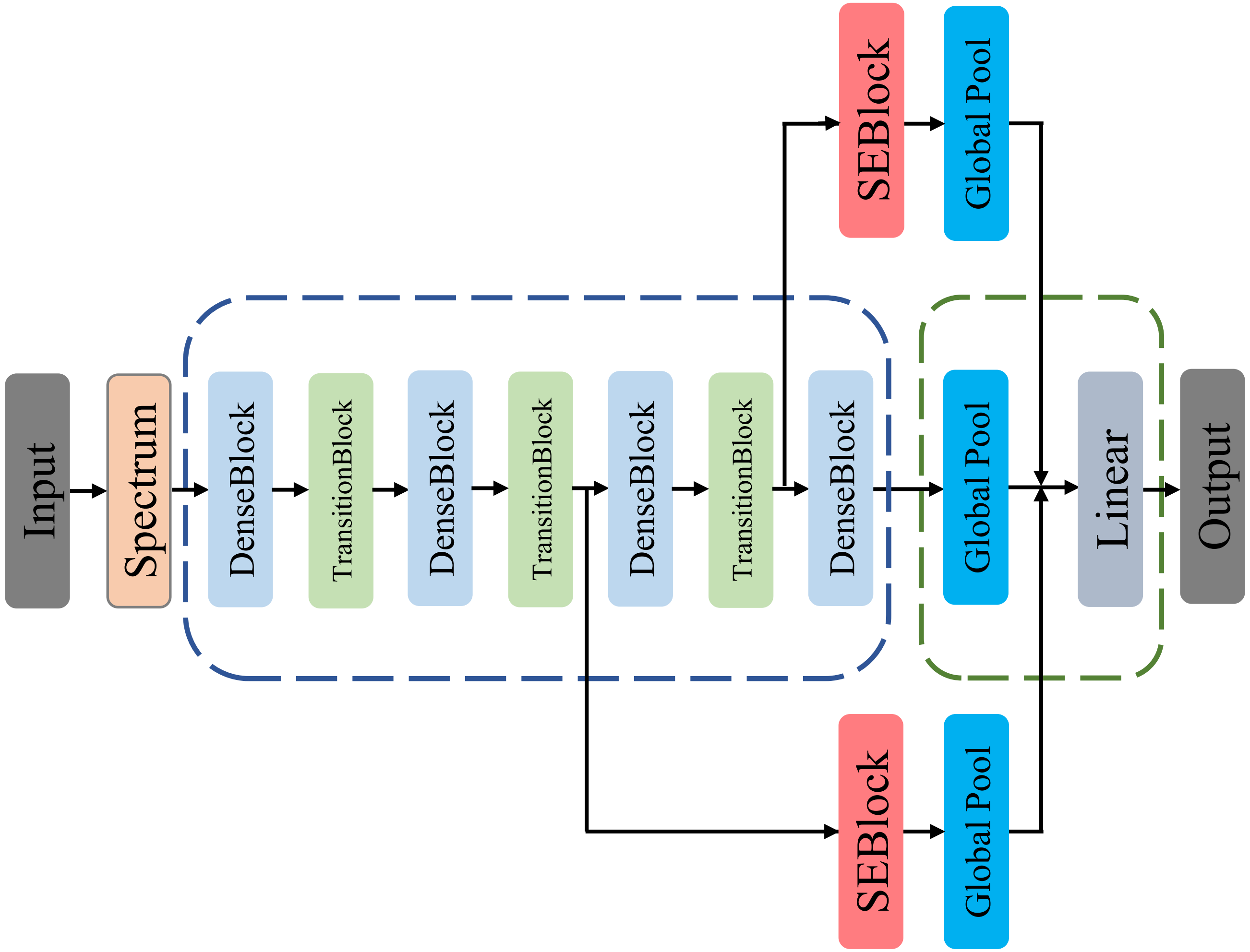
Model Result

Model Name	Parameters	Input Feature	ESC50
EnvNet2	101.25M	wav	84.70
Proposed1	7.03M		87.85
Proposed2	7.46M (+6.12%)		88.05 (+0.20)

Rethink Feature Extraction



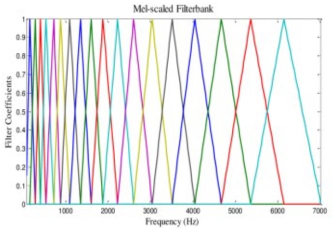
Proposed3 - Input Spectrum

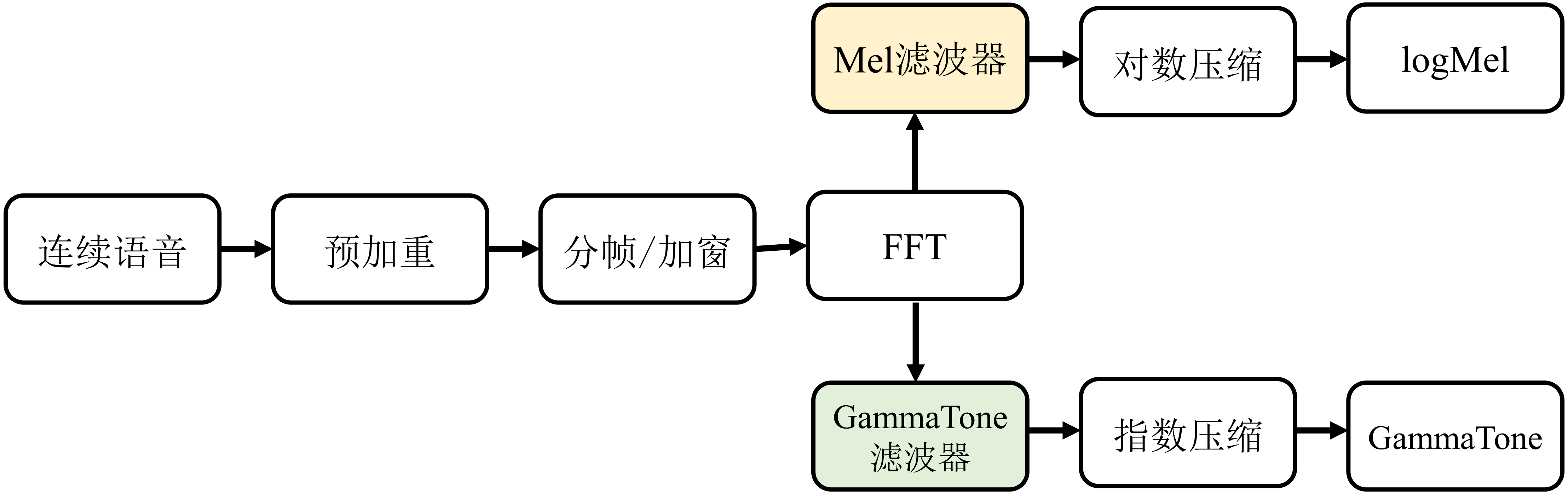


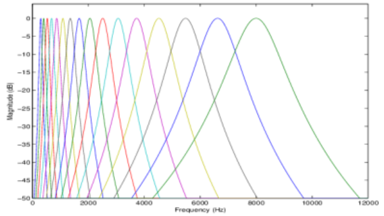
Spectrum Select

- 1 logMel
- 2 GammaTone
- 3 Constant Q-transform
- 4 logMel + GammaTone
- 5 logMel + First Derivative + Second Derivative

Spectrum Extraction







Model Result

Model Name	Parameters	Input Feature	ESC50
EnvNet2	101.25M	wav	84.70
Proposed1	7.03M		87.85
Proposed2	7.46M		88.05
Proposed3	7.43M (-0.4%)	logMel	91.05 (+3.00)
		GammaTone	87.25 (fold1) [1]
		logMel+GammaTone	89.75 (+1.7)
		logMel + Derivative	89.75(+1.7)

[1]　fold1结果：EnvNet2 80.75、Proposed1 86.75、Proposed2 89.00、Proposed3(logMel) 90.50

Rethink Conv Kernel

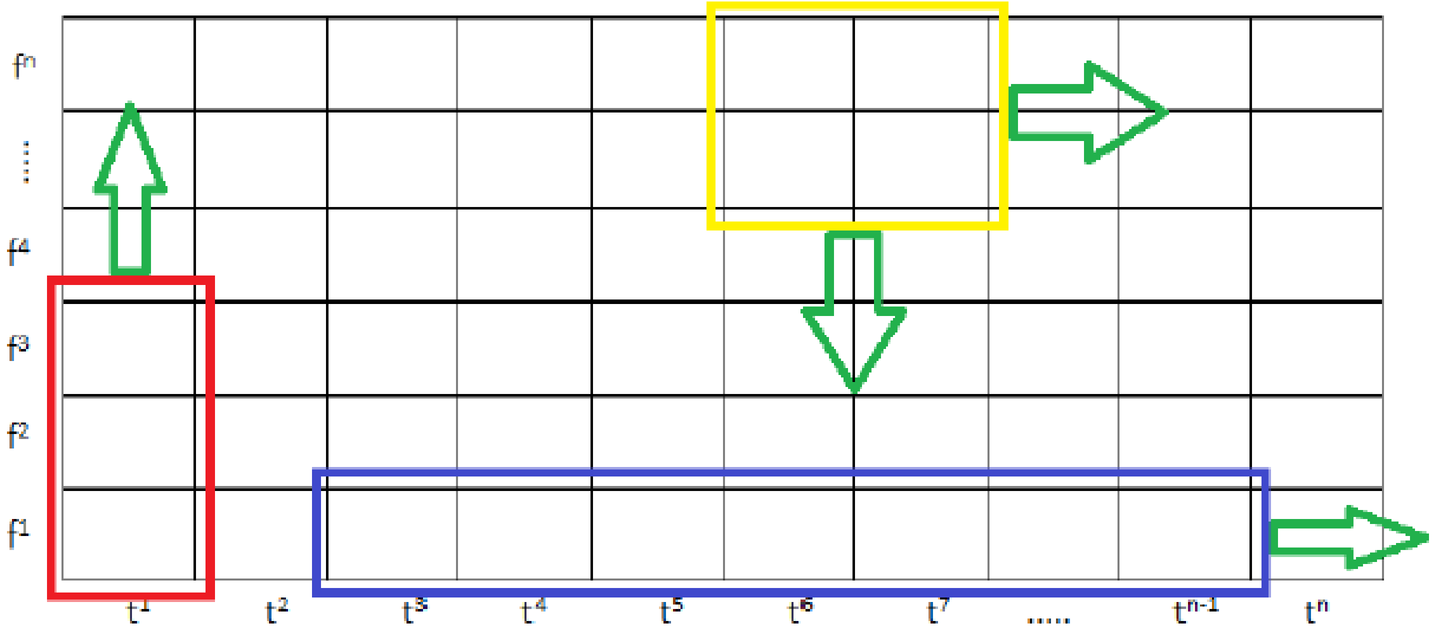


Fig. 2: Separable Convolutions working in the time and feature domains vs Standard Convolutions

Proposed4 - Separable Conv

Layers	DenseNet-121
DenseBlock1	[1x1 Conv 3x3 Conv] x 6
TransitionBlock1	1x1 Conv 2x2 AvgPool
DenseBlock2	[1x1 Conv 3x3 Conv] x 12
TransitionBlock2	1x1 Conv 2x2 AvgPool
DenseBlock3	[1x1 Conv 3x3 Conv] x 24
TransitionBlock3	1x1 Conv 2x2 AvgPool
DenseBlock4	[1x1 Conv 3x3 Conv] x 16

DenseBlock1 ⇒

1x1 Conv
1x3 Conv x 3

1x1 Conv
3x1 Conv x 3

DenseBlock2 ⇒

1x1 Conv
1x3 Conv x 6

1x1 Conv
3x1 Conv x 6

Model Result

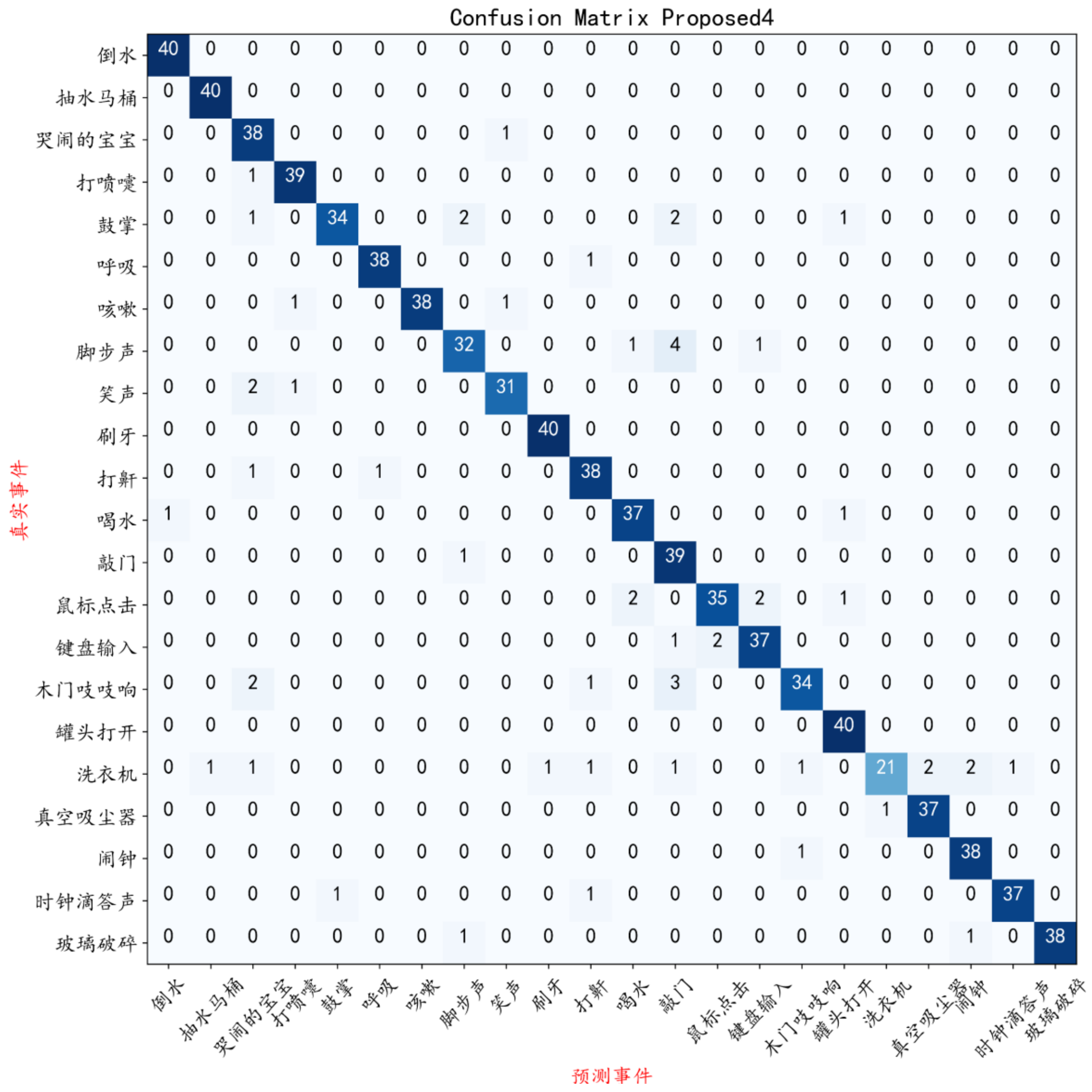
Model Name	Parameters	Input Feature	ESC50
EnvNet2	101.25M	wav	84.70
Proposed1	7.03M		87.85
Proposed2	7.46M		88.05
Proposed3	7.43M	logMel	91.05
Proposed4	6.63M (-10.77%)	logMel	91.35 (+0.3)
Simple-Proposed4 [1]	1.85M (-75.1%)		89.55 (-1.5)

[1] DenesNet(growth_rate=16, block_config=(6,12,24,16)) 通道参数减半

Previous State-of-the-art Models vs Proposed

Model Name	Input Feature	ESC50	Urbansound8K official	Urbansound8K unofficial
Human	-	81.30	-	-
WaveMSNet(2018)	wav	79.10	-	-
EnvNet(2017)		74.10	71.10	-
EnvNet2(2018)		84.70	78.30	-
Piczak-CNN (2015)	Mel-Spectrum	64.50	73.70	-
TFNet(2019)		87.70	-	88.50
ESResNet(2020)		83.15	82.76	**96.83**
Piczak-CNN (2017)	GammaTone-Spectrum	81.95	-	88.02
VGG-like-CNN (2019)		86.50	-	-
AlexNet(2017)	混合特征	65.00	-	92.00
GoogleNet (2017)		73.00	-	93.00
VGG-like-CNN (2018)		83.90	**83.70**	-
Separable CNN (2019)		**89.75**	-	91.75
Proposed4	Mel-Spectrum	**91.35**	**83.70**	**98.67**

Confusion Matrix Proposed4



Conclusion

- Feature Select： 从Waveform到Spectrum, 尤其表现优异logMel；
- Model Structure: 从DenseNet到Multi-Scale、SEBlock再到Separable Conv，模型参数降低的同时，准确率也显著提升；

Future Work

- Feature Fusion: 继续从数据类型、特征分布以及模型结构上调研和探索当前模型特征融合的提升。

- Ensemble Model: 多模型集成提升准确率，多种组合方式共同表决结果。

- Dynamic Conv: 一种动态卷积机制，它有助于提升模型的特征表达能力, 无需提升网络深度与宽度。

Acknowledgment

References

ResNet or DenseNet

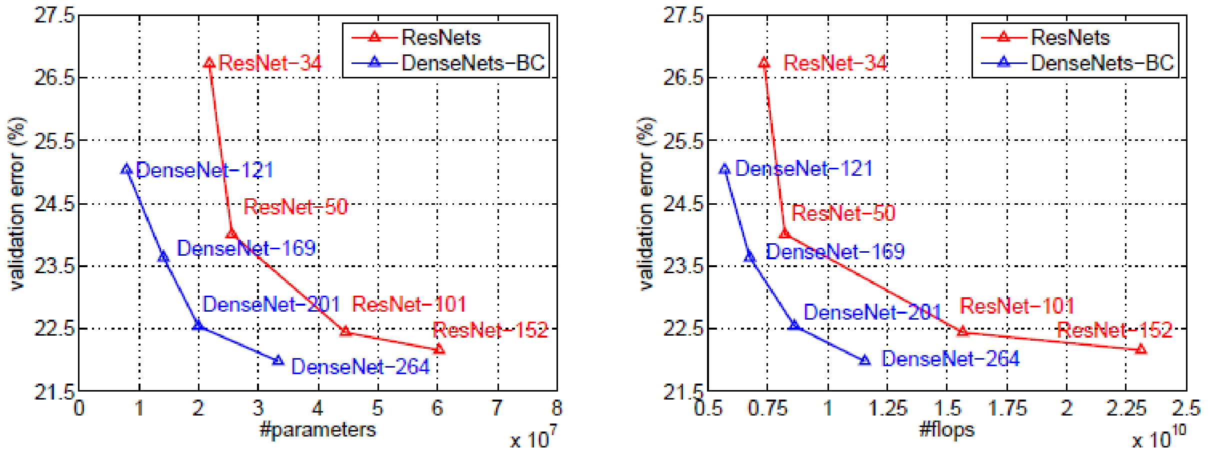


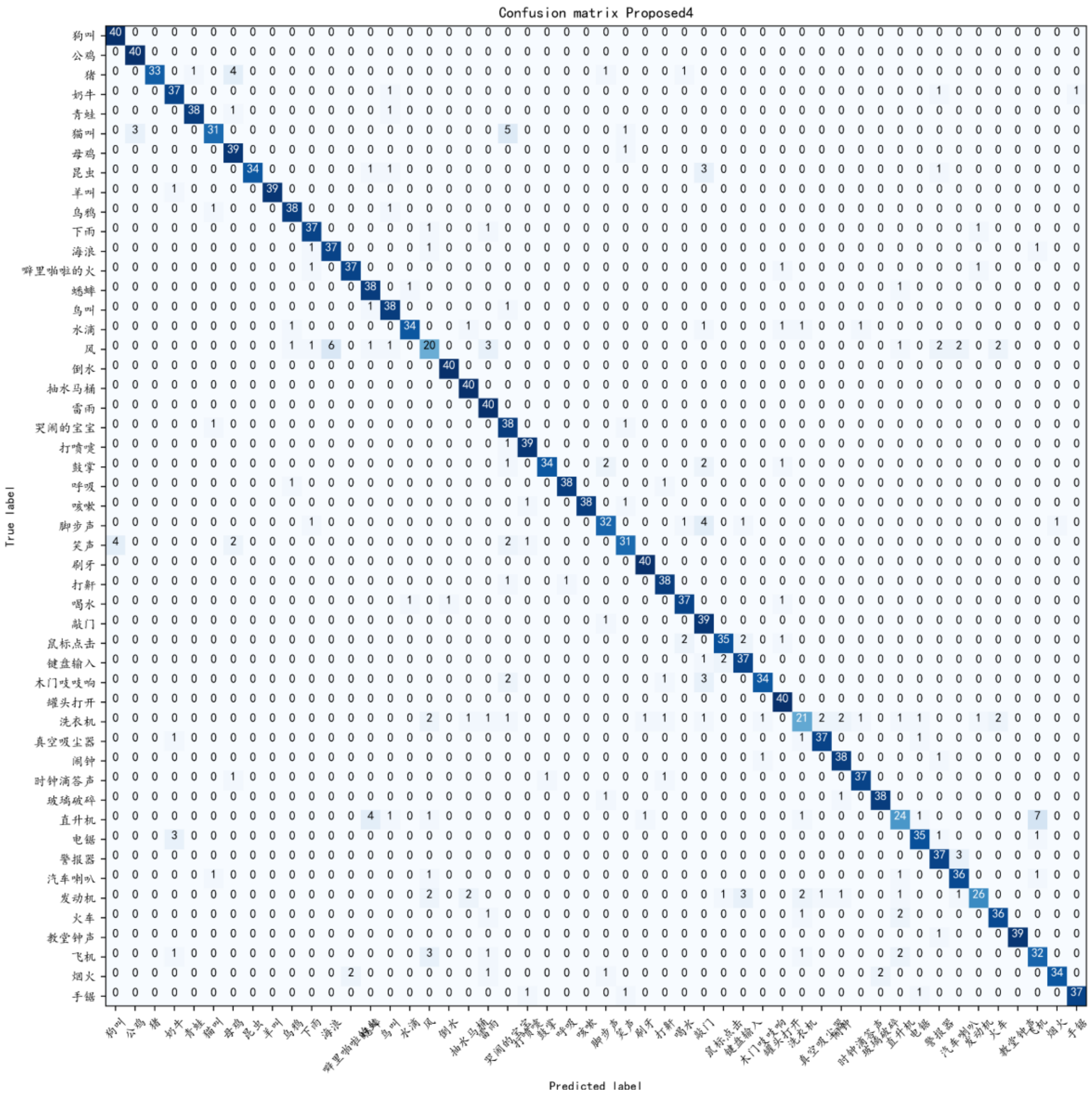
Figure 3: Comparison of the DenseNets and ResNets top-1 error rates (single-crop testing) on the ImageNet validation dataset as a function of learned parameters (*left*) and FLOPs during test-time (*right*).

SENet

TABLE 2
Single-crop error rates (%) on the ImageNet validation set and complexity comparisons. The *original* column refers to the results reported in the original papers (the results of ResNets are obtained from the website: https://github.com/Kaiminghe/deep-residual-networks). To enable a fair comparison, we re-train the baseline models and report the scores in the *re-implementation* column. The *SENet* column refers to the corresponding architectures in which SE blocks have been added. The numbers in brackets denote the performance improvement over the re-implemented baselines. † indicates that the model has been evaluated on the non-blacklisted subset of the validation set (this is discussed in more detail in [21]), which may slightly improve results. VGG-16 and SE-VGG-16 are trained with batch normalization.

	original		re-implementation			SENet		
	top-1 err.	top-5 err.	top-1 err.	top-5 err.	GFLOPs	top-1 err.	top-5 err.	GFLOPs
ResNet-50 [13]	24.7	7.8	24.80	7.48	3.86	$23.29_{(1.51)}$	$6.62_{(0.86)}$	3.87
ResNet-101 [13]	23.6	7.1	23.17	6.52	7.58	$22.38_{(0.79)}$	$6.07_{(0.45)}$	7.60
ResNet-152 [13]	23.0	6.7	22.42	6.34	11.30	$21.57_{(0.85)}$	$5.73_{(0.61)}$	11.32
ResNeXt-50 [19]	22.2	-	22.11	5.90	4.24	$21.10_{(1.01)}$	$5.49_{(0.41)}$	4.25
ResNeXt-101 [19]	21.2	5.6	21.18	5.57	7.99	$20.70_{(0.48)}$	$5.01_{(0.56)}$	8.00
VGG-16 [11]	-	-	27.02	8.81	15.47	$25.22_{(1.80)}$	$7.70_{(1.11)}$	15.48
BN-Inception [6]	25.2	7.82	25.38	7.89	2.03	$24.23_{(1.15)}$	$7.14_{(0.75)}$	2.04
Inception-ResNet-v2 [21]	$19.9^{\dagger}$	$4.9^{\dagger}$	20.37	5.21	11.75	$19.80_{(0.57)}$	$4.79_{(0.42)}$	11.76

Confusion Matrix Proposed4



Mix-up Data Augmentation

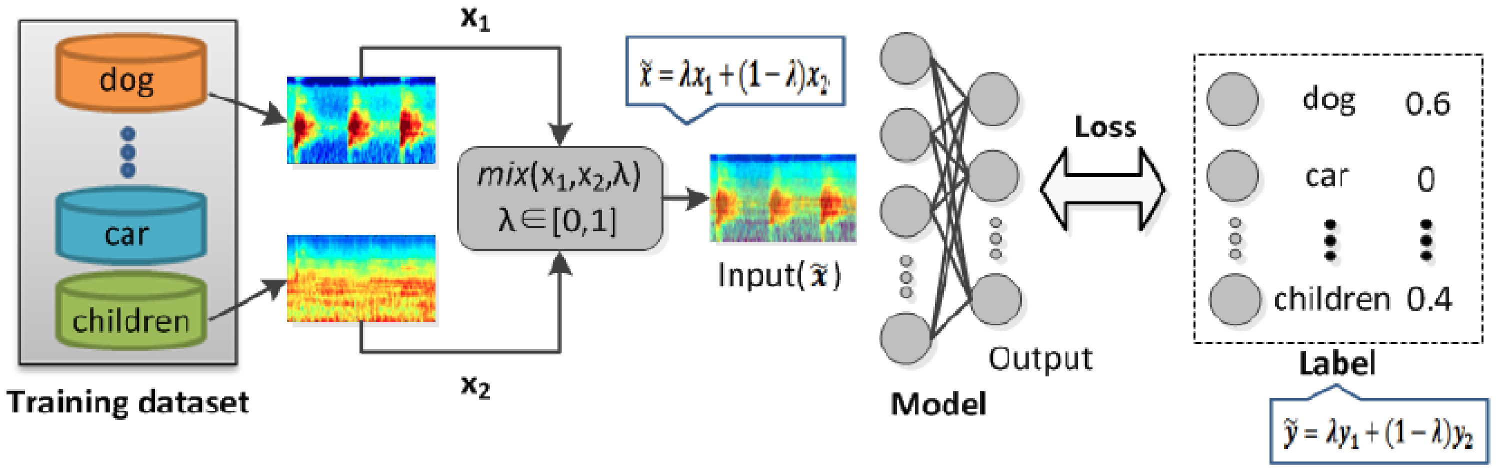


Fig. 1. Pipeline of mixup. Every training sample is created by mixing two examples randomly selected from original training dataset. We use the mixed sound to train the model and the train target is the mixing ratio.