AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

HONG IK UNIVERSITY

SINGAPORE INSTITUTE OF TECHNOLOGY

A Low-Complexity Deep Learning Framework For Acoustic Scene Classification

Presenter: **Lam Pham** from **AIT**

IDSC 2021

International Data Science Conference 2021

Agenda

1. Task Definition & Challenges

2. The Framework Proposed

3. Results & Discussion

Agenda

1. Task Definition & Challenges

2. The Framework Proposed

3. Results & Discussion

Task Definition & Challenges

VS.

Machine Hearing

Acoustic Scene Classifier (ASC)

Acoustic Event Detection (AED)

Human and Machine Hearing is a book for people who want to understand how the auditory system and the brain process sound, how to encapsulate aspects of our hearing knowledge in computer algorithms, and how to combine the algorithms into a machine that simulates the role of hearing in some aspect of everyday life—such as listening to the melody of a song or talking to a friend in a noisy restaurant. This is what Dick Lyon means by "Machine Hearing."

F.Lyon Richard, Human and Machine Hearing, p. xi. Cambridge University Press, 2017.

Task Definition & Challenges

VS.

Machine Hearing

Acoustic Scene Classifier (ASC)

Acoustic Event Detection (AED)

Human and Machine Hearing is a book for people who want to understand how the auditory system and the brain process sound, how to encapsulate aspects of our hearing knowledge in computer algorithms, and how to combine the algorithms into a machine that simulates the role of hearing in some aspect of everyday life—such as listening to the melody of a song or talking to a friend in a noisy restaurant. This is what Dick Lyon means by "Machine Hearing."

F.Lyon Richard, Human and Machine Hearing, p. xi. Cambridge University Press, 2017.

Task Definition & Challenges

Acoustic Scene Classifier (ASC)

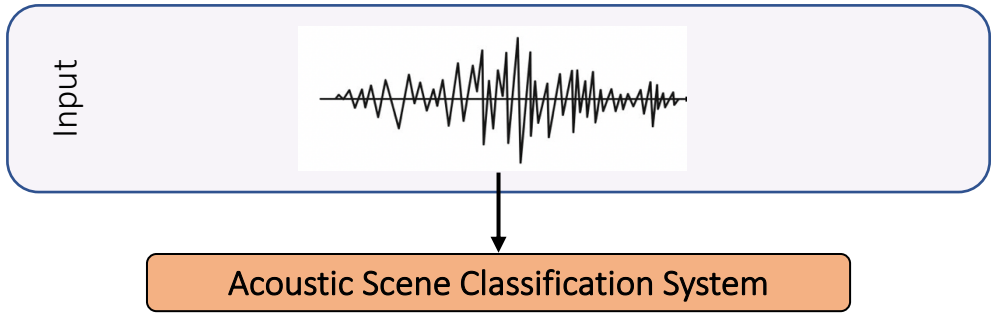
Task Definition & Challenges

Acoustic Scene Classifier (ASC)

Input

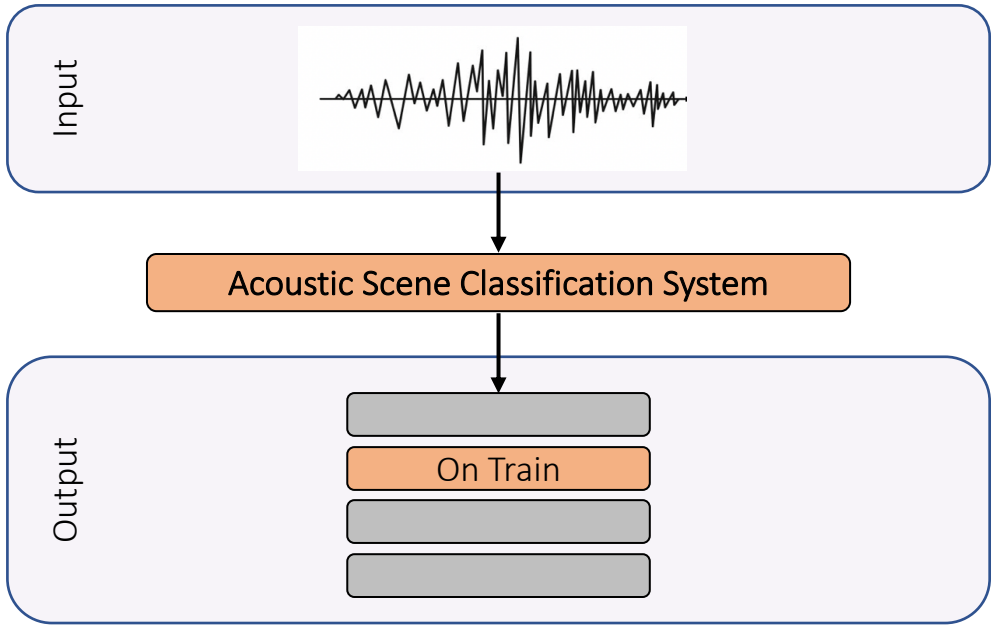
Task Definition & Challenges

Acoustic Scene Classifier (ASC)



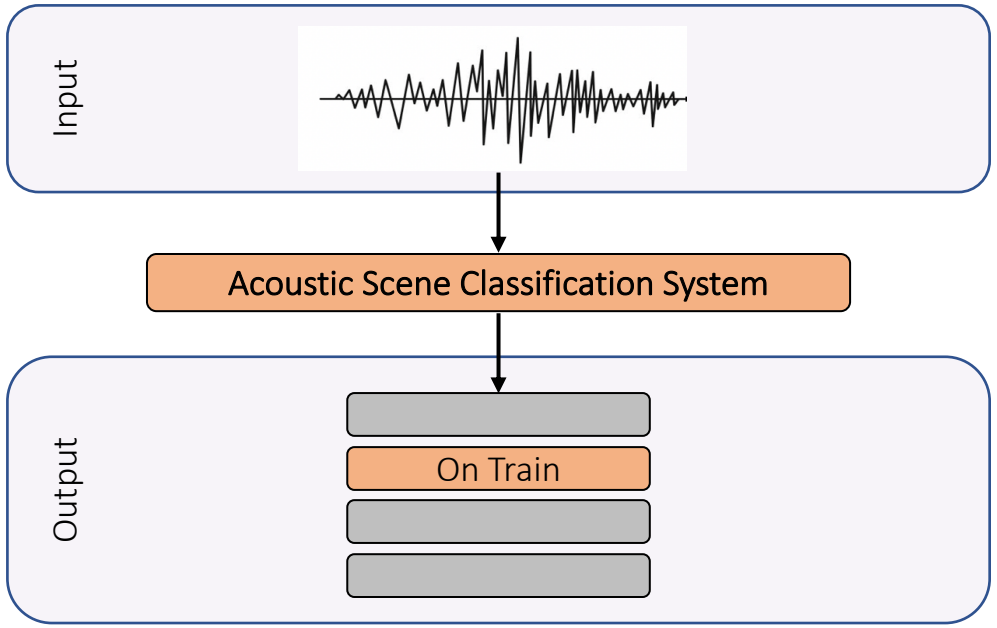
Task Definition & Challenges

Acoustic Scene Classifier (ASC)



Task Definition & Challenges

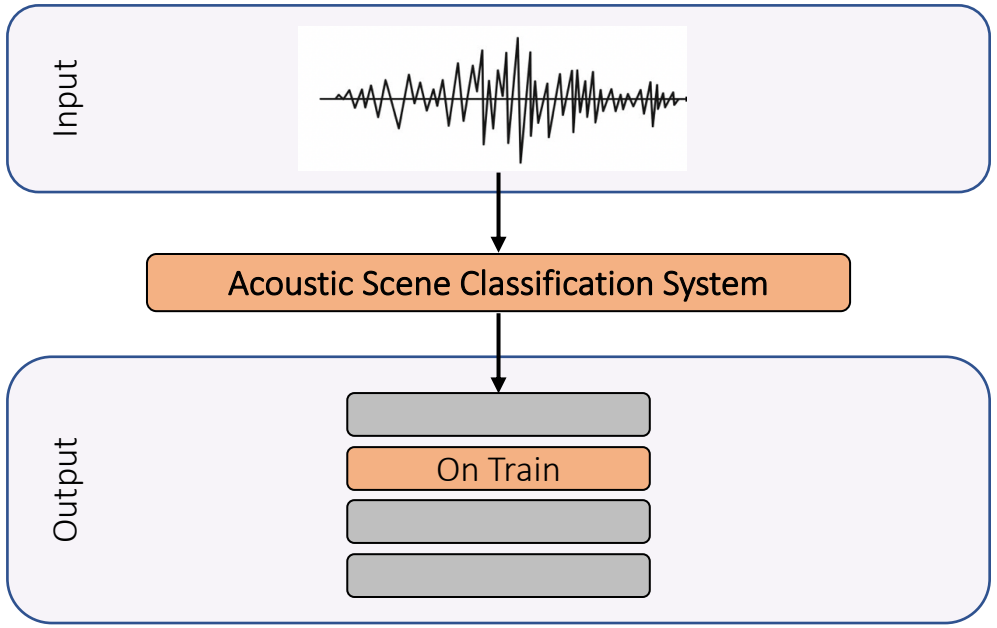
Acoustic Scene Classifier (ASC)



→ Basically, this is supervised classification

Task Definition & Challenges

Acoustic Scene Classifier (ASC)

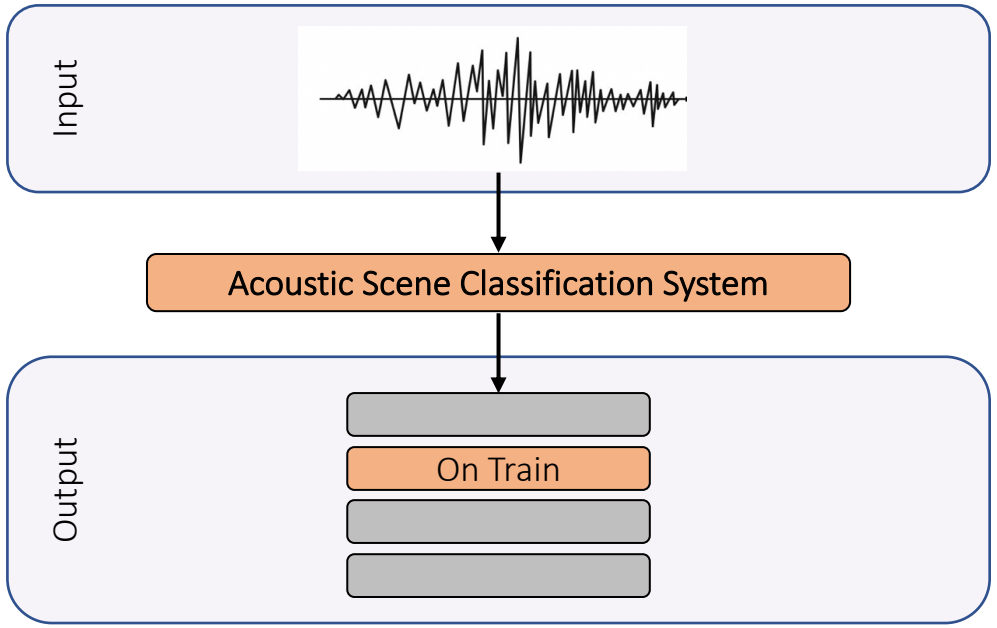


→ Basically, this is supervised classification

State of the art

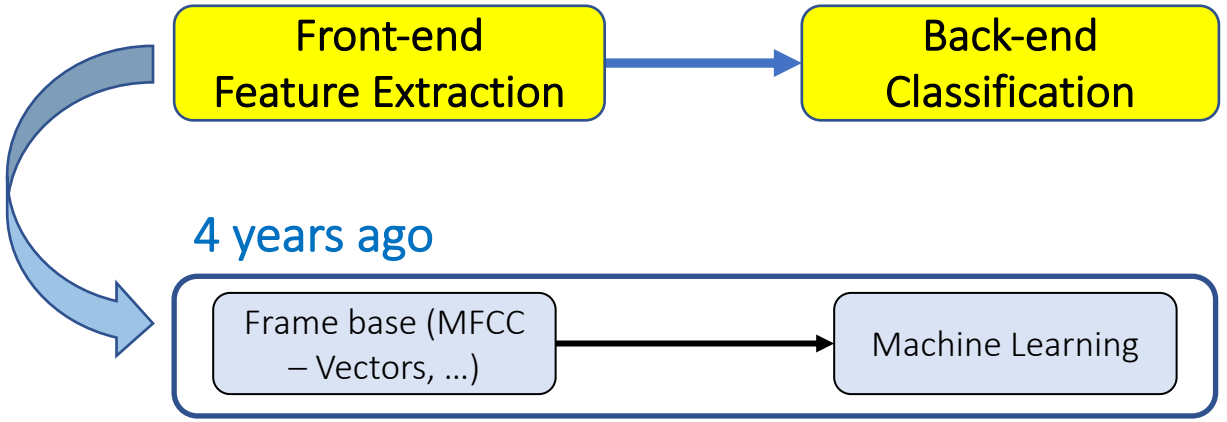
Task Definition & Challenges

Acoustic Scene Classifier (ASC)



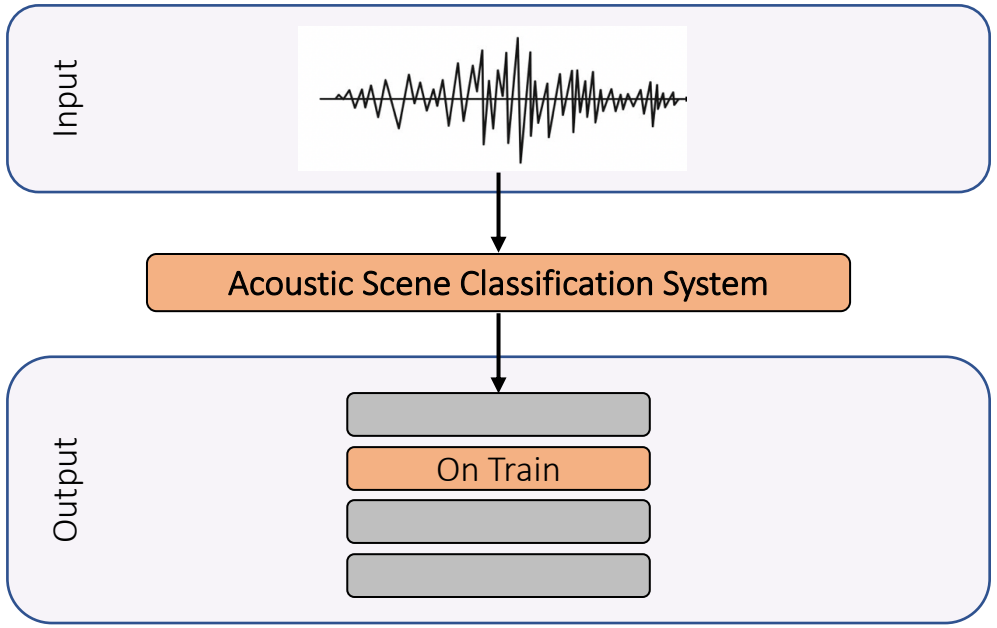
→ Basically, this is supervised classification

State of the art



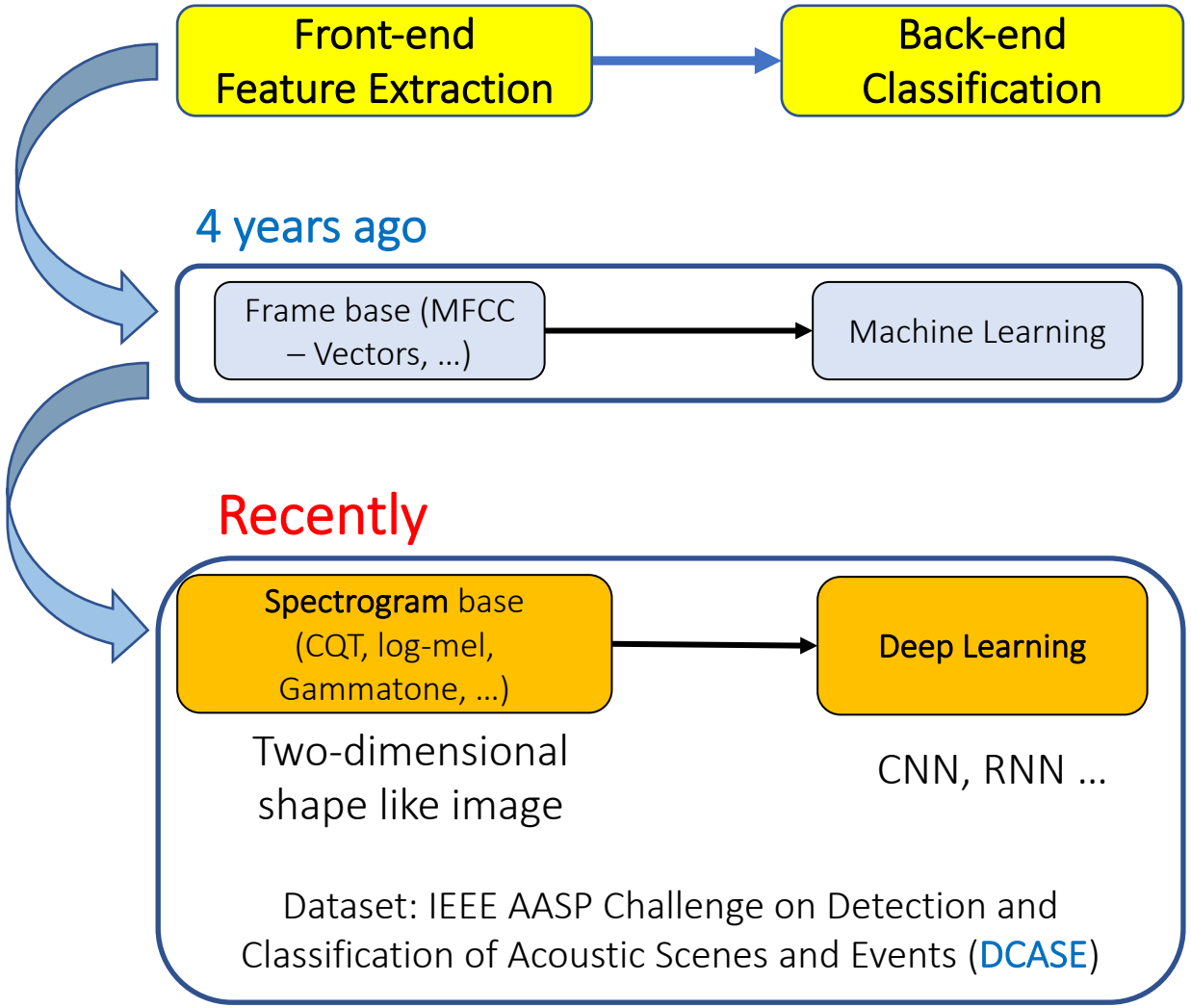
Task Definition & Challenges

Acoustic Scene Classifier (ASC)



→ Basically, this is supervised classification

State of the art



Task Definition & Challenges

Challenges

Task Definition & Challenges

Challenges

- Diverse background noise
- SNR varies
- Non structure as speech

...

Task Definition & Challenges

Challenges

Solutions

- Diverse background noise
- SNR varies
- Non structure as speech

…

Task Definition & Challenges

Challenges

- Diverse background noise
- SNR varies
- Non structure as speech

…

Solutions

- Multiple input features
--> tackle lacking of input

- Ensemble of classifiers
--> create general model

- Complex classifiers
--> learn feature deeply

Task Definition & Challenges

Challenges

- Diverse background noise
- SNR varies
- Non structure as speech

…

Solutions

- Multiple input features
--> tackle lacking of input

- Ensemble of classifiers
--> create general model

- Complex classifiers
--> learn feature deeply

- Computation cost
- Model complexity
- Overfitting on one dataset
- Challenges for edge-device based applications

Task Definition & Challenges

Challenges

- Diverse background noise
- SNR varies
- Non structure as speech

…

Solutions

- Multiple input features
  --> tackle lacking of input

- Ensemble of classifiers
  --> create general model

- Complex classifiers
  --> learn feature deeply

- Computation cost
- Model complexity
- Overfitting on one dataset
- Challenges for edge-device based applications

Performance vs. Complexity

Agenda

1. Motivation & Dataset

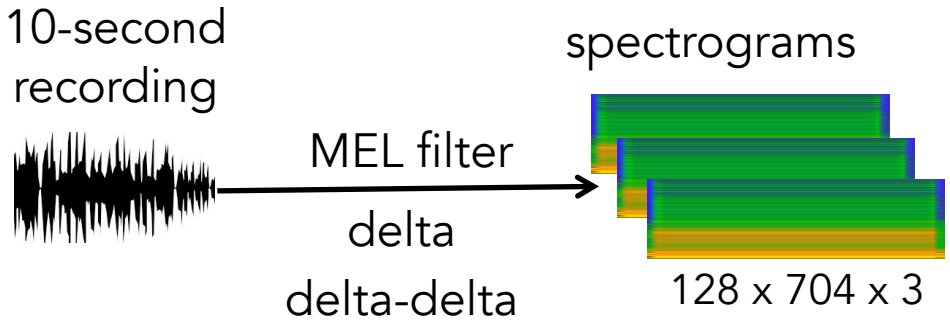
2. The Framework Proposed

3. Results & Discussion

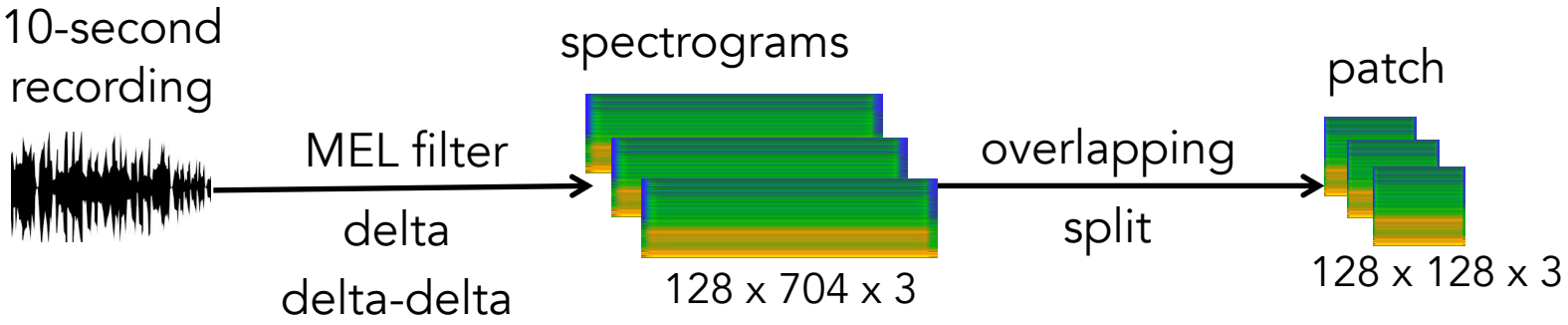
The Framework Proposed - *Baseline*

10-second recording

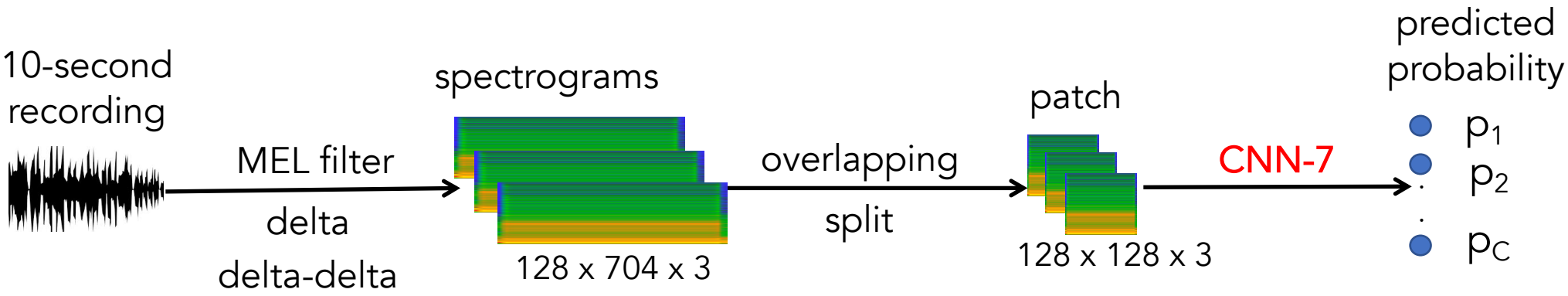
The Framework Proposed - *Baseline*



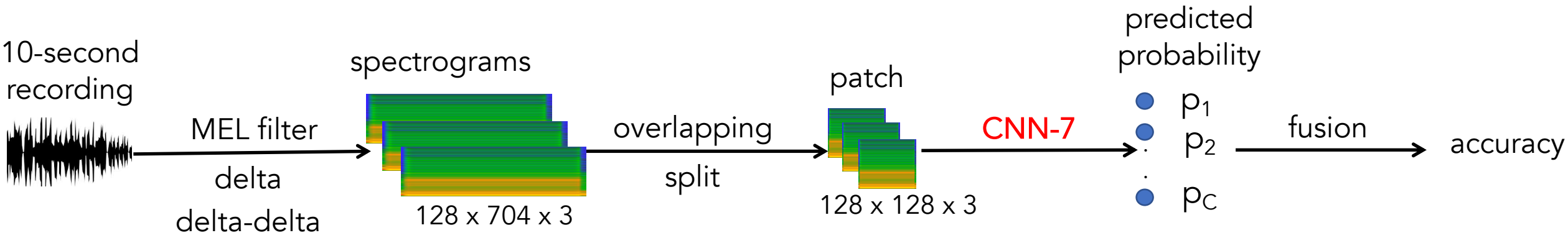
The Framework Proposed - *Baseline*



The Framework Proposed - *Baseline*



The Framework Proposed - *Baseline*



The Framework Proposed - *Baseline*

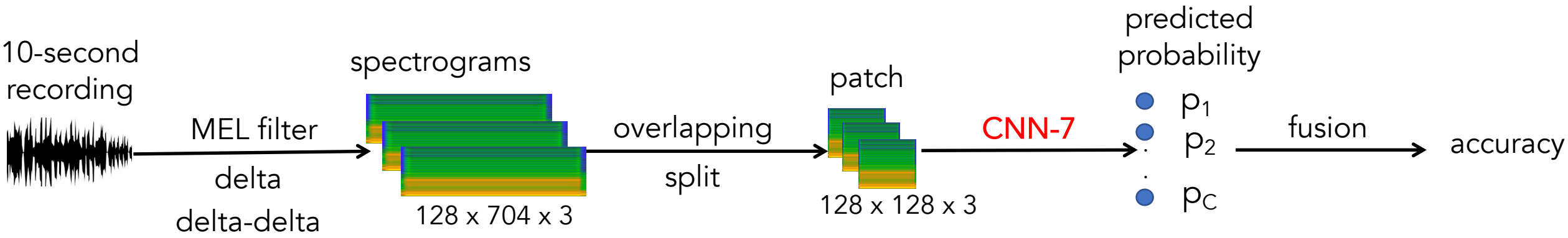


TABLE I

THE CNN-7 NETWORK ARCHITECTURE BASELINE (INPUT PATCH OF 128×128×3)

Network architecture	**Output**
BN - Convolution ([3×3] @$C_{out}1 = 32$) - ReLU - BN - Dropout (10%)	128×128×32
BN - Convolution ([3×3] @$C_{out}2 = 32$) - ReLU - BN - AP [2×2] - Dropout (10%)	64×64×32
BN - Convolution ([3×3] @$C_{out}3 = 64$) - ReLU - BN - Dropout (10%)	64×64×64
BN - Convolution ([3×3] @$C_{out}4 = 64$) - ReLU - BN - AP [2×2] - Dropout (10%)	32×32×64
BN - Convolution ([3×3] @$C_{out}5 = 128$) - ReLU - BN - AP [2×2] - Dropout (10%)	16×16×128
BN - Convolution ([3×3] @$C_{out}6 = 128$) - ReLU - BN - GAP - Dropout (10%)	128
FC - Softmax	$C = 10$

The Framework Proposed - *Baseline*

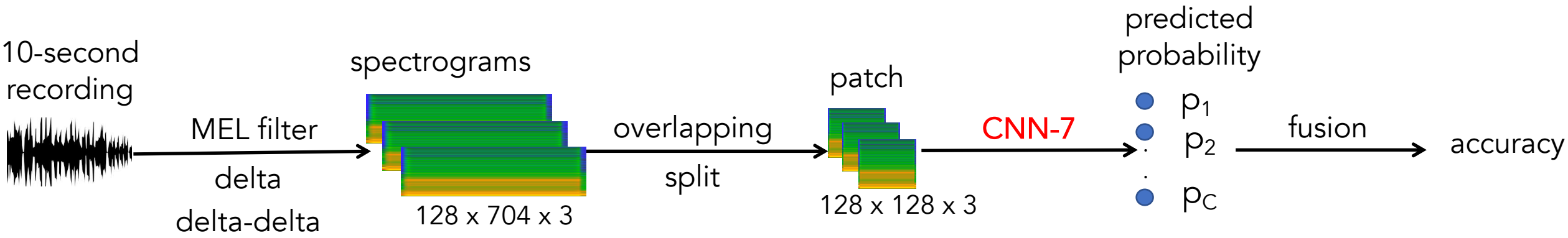


TABLE I

THE CNN-7 NETWORK ARCHITECTURE BASELINE (INPUT PATCH OF 128×128×3)

Network architecture	Output
BN - Convolution ([3×3] @$C_{out}1 = 32$) - ReLU - BN - Dropout (10%)	128×128×32
BN - Convolution ([3×3] @$C_{out}2 = 32$) - ReLU - BN - AP [2×2] - Dropout (10%)	64×64×32
BN - Convolution ([3×3] @$C_{out}3 = 64$) - ReLU - BN - Dropout (10%)	64×64×64
BN - Convolution ([3×3] @$C_{out}4 = 64$) - ReLU - BN - AP [2×2] - Dropout (10%)	32×32×64
BN - Convolution ([3×3] @$C_{out}5 = 128$) - ReLU - BN - AP [2×2] - Dropout (10%)	16×16×128
BN - Convolution ([3×3] @$C_{out}6 = 128$) - ReLU - BN - GAP - Dropout (10%)	128
FC - Softmax	$C = 10$

Our proposed single CNN-7 architecture reports a complexity of 1,129 MB for non-zero parameters with using 32 bits for representing one parameter

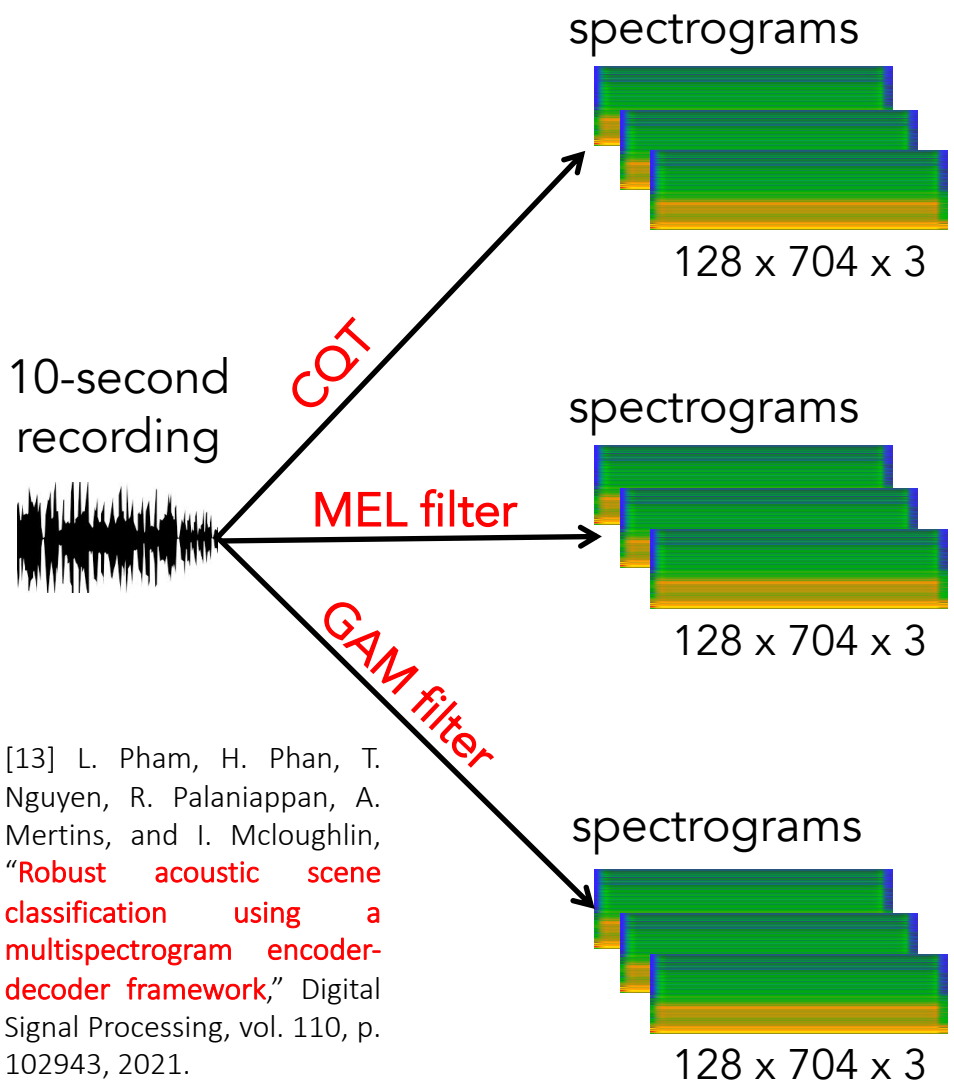
The Framework Proposed - *Architecture*

Multiple Spectrograms

Improve performance, but the model complexity (3 individual CNN-7) increases to 3 times.

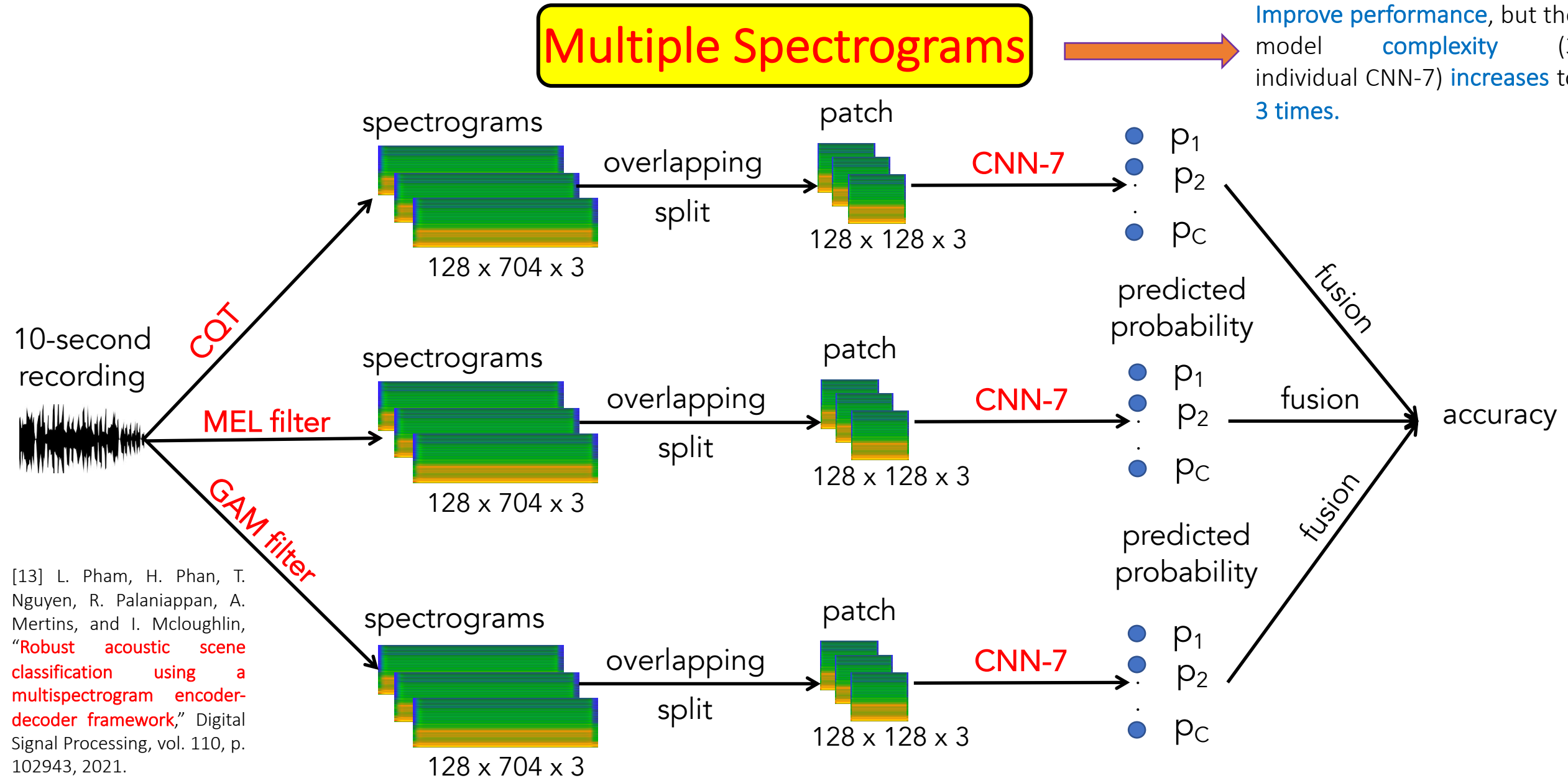
The Framework Proposed - *Architecture*

Multiple Spectrograms → Improve performance, but the model complexity (3 individual CNN-7) increases to 3 times.



[13] L. Pham, H. Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. Mcloughlin, "Robust acoustic scene classification using a multispectrogram encoder-decoder framework," Digital Signal Processing, vol. 110, p. 102943, 2021.

The Framework Proposed - *Architecture*



[13] L. Pham, H. Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. Mcloughlin, "Robust acoustic scene classification using a multispectrogram encoder-decoder framework," Digital Signal Processing, vol. 110, p. 102943, 2021.

The Framework Proposed - *Architecture*

Model Decompression

The model complexity
(individual CNN-7 using CR & DC)
is reduced to nearly 1/34 times.

The Framework Proposed - *Architecture*

Model Decompression → The model complexity (individual CNN-7 using CR & DC) is reduced to nearly 1/34 times.

1. Channel Restriction (CR)

TABLE I
THE CNN-7 NETWORK ARCHITECTURE BASELINE (INPUT PATCH OF 128×128×3)

Network architecture	Output
BN - Convolution ([3×3]@$C_{out}1 = 32$) - ReLU - BN - Dropout (10%)	128×128×32
BN - Convolution ([3×3]@$C_{out}2 = 32$) - ReLU - BN - AP [2×2] - Dropout (10%)	64×64×32
BN - Convolution ([3×3]@$C_{out}3 = 64$) - ReLU - BN - Dropout (10%)	64×64×64
BN - Convolution ([3×3]@$C_{out}4 = 64$) - ReLU - BN - AP [2×2] - Dropout (10%)	32×32×64
BN - Convolution ([3×3]@$C_{out}5 = 128$) - ReLU - BN - AP [2×2] - Dropout (10%)	16×16×128
BN - Convolution ([3×3]@$C_{out}6 = 128$) - ReLU - BN - GAP - Dropout (10%)	128
FC - Softmax	$C = 10$

[30] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer, "Lowcomplexity models for acoustic scene classification based on receptive field regularization and frequency damping," arXiv preprint arXiv:2011.02955, 2020.

The Framework Proposed - *Architecture*

Model Decompression → The model complexity (individual CNN-7 using CR & DC) is reduced to nearly 1/34 times.

1. Channel Restriction (CR)

Restrict the number of channels: Cout1 from 32 to 16, Cout3 and Cout4 from 64 to 32, Cout5 and Cout6 from 128 to 64.

→ reduce an individual CNN-7 complexity to 313 KB that nearly equals to 1/4 of the original size (1,129 MB).

TABLE I
THE CNN-7 NETWORK ARCHITECTURE BASELINE (INPUT PATCH OF 128×128×3)

Network architecture	Output
BN - Convolution ([3×3] @$C_{out}1 = 32$) - ReLU - BN - Dropout (10%)	128×128×32
BN - Convolution ([3×3] @$C_{out}2 = 32$) - ReLU - BN - AP [2×2] - Dropout (10%)	64×64×32
BN - Convolution ([3×3] @$C_{out}3 = 64$) - ReLU - BN - Dropout (10%)	64×64×64
BN - Convolution ([3×3] @$C_{out}4 = 64$) - ReLU - BN - AP [2×2] - Dropout (10%)	32×32×64
BN - Convolution ([3×3] @$C_{out}5 = 128$) - ReLU - BN - AP [2×2] - Dropout (10%)	16×16×128
BN - Convolution ([3×3] @$C_{out}6 = 128$) - ReLU - BN - GAP - Dropout (10%)	128
FC - Softmax	$C = 10$

[30] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer, "Lowcomplexity models for acoustic scene classification based on receptive field regularization and frequency damping," arXiv preprint arXiv:2011.02955, 2020.

The Framework Proposed - *Architecture*

Model Decompression → The model complexity (individual CNN-7 using CR & DC) is reduced to nearly 1/34 times.

1. Channel Restriction (CR)

Restrict the number of channels: Cout1 from 32 to 16, Cout3 and Cout4 from 64 to 32, Cout5 and Cout6 from 128 to 64.

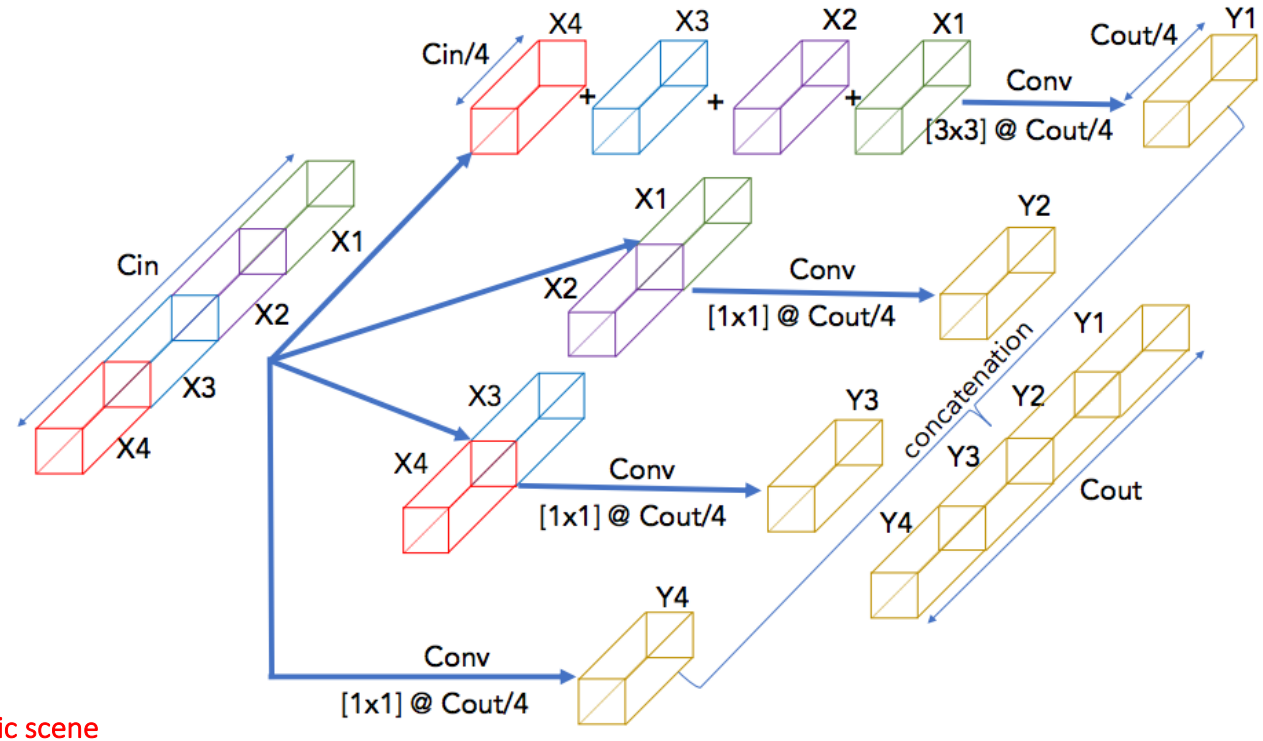
→ reduce an individual CNN-7 complexity to 313 KB that nearly equals to 1/4 of the original size (1,129 MB).

TABLE I
THE CNN-7 NETWORK ARCHITECTURE BASELINE (INPUT PATCH OF 128×128×3)

Network architecture	Output
BN - Convolution ([3×3]@$C_{out}1 = 32$) - ReLU - BN - Dropout (10%)	128×128×32
BN - Convolution ([3×3]@$C_{out}2 = 32$) - ReLU - BN - AP [2×2] - Dropout (10%)	64×64×32
BN - Convolution ([3×3]@$C_{out}3 = 64$) - ReLU - BN - Dropout (10%)	64×64×64
BN - Convolution ([3×3]@$C_{out}4 = 64$) - ReLU - BN - AP [2×2] - Dropout (10%)	32×32×64
BN - Convolution ([3×3]@$C_{out}5 = 128$) - ReLU - BN - AP [2×2] - Dropout (10%)	16×16×128
BN - Convolution ([3×3]@$C_{out}6 = 128$) - ReLU - BN - GAP - Dropout (10%)	128
FC - Softmax	$C = 10$

2. Decomposed convolution (DC)

The individual CNN-7 complexity is reduced to nearly 1/8.5 of the original size (1,129 MB).



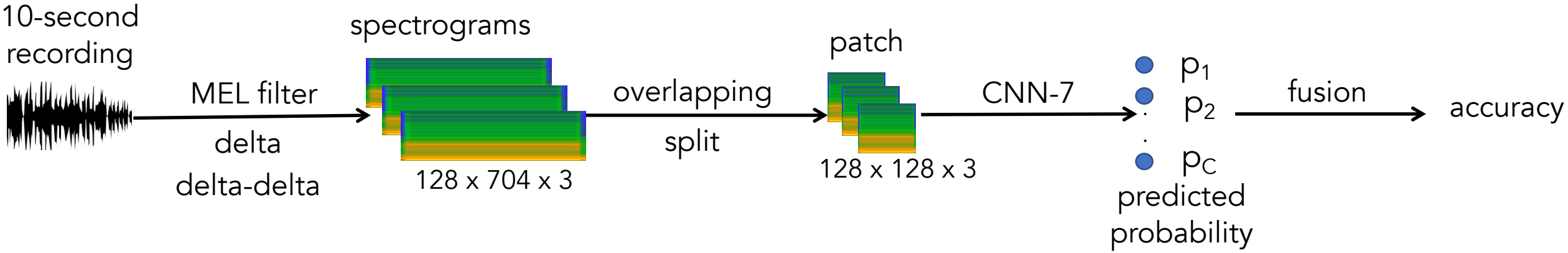
[30] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer, "Lowcomplexity models for acoustic scene classification based on receptive field regularization and frequency damping," arXiv preprint arXiv:2011.02955, 2020.

The Framework Proposed - *Architecture*

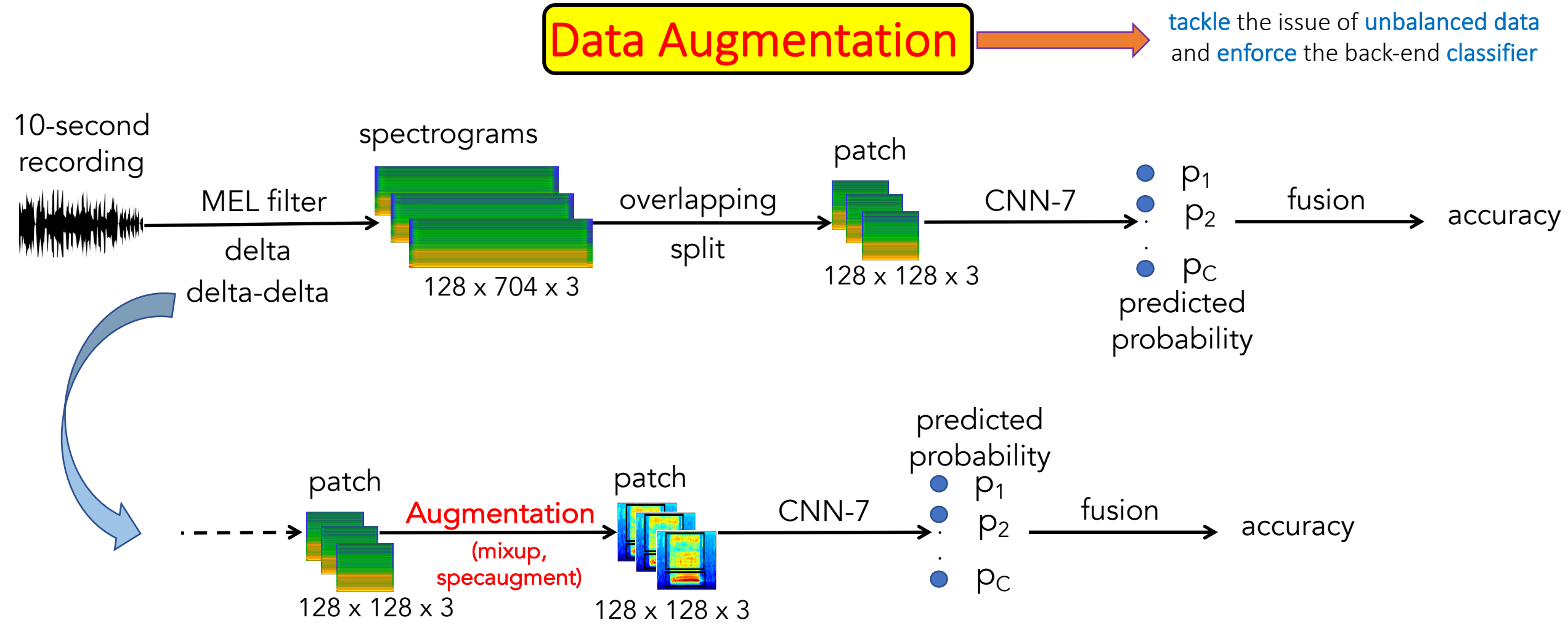
Data Augmentation → tackle the issue of unbalanced data and enforce the back-end classifier

The Framework Proposed - *Architecture*

Data Augmentation → tackle the issue of unbalanced data and enforce the back-end classifier



The Framework Proposed - *Architecture*



[36] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in Pacific Rim Conference on Multimedia, 2018, pp. 14–23.

[38] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.

The Framework Proposed – *Experimental Setting*

TABLE II
THE NUMBER OF 10-SECOND AUDIO RECORDINGS CORRESPONDING TO EACH SCENE CATEGORIES IN THE TRAIN. AND EVAL. SUBSETS SEPARATED FROM THE DCASE 2021 TASK 1A DEVELOPMENT DATASET [35], AND THE EVALUATION DATASET [32].

Category	Train. Subset	Eval. Subset	Evaluation
Airport	1393	296	-
Bus	1400	297	-
Metro	1382	297	-
Metro Station	1380	297	-
Park	1429	297	-
Public square	1427	297	-
Shopping mall	1373	297	-
Street pedestrian	1386	297	-
Street traffic	1413	297	-
Tram	1379	296	-
Total	13962 (≈38.79 hours)	2968 (≈8.25 hours)	7920 (22 hours)

* The **Development dataset (Train. & Eval. Subsets):** The dataset in slightly unbalanced, recorded from 12 large European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna.
The audio recordings were recorded from 3 different devices namely A (10215 recordings), B (749 recordings), C (748 recordings). Additionally, synthetic data for 11 mobile devices was created based on the original recordings, referred to as S1 (750 recordings), S2 (750 recordings), S3 (750 recordings), S4 (750 recordings), S5 (750 recordings), and S6 (750 recordings).

* The **Evaluation dataset without labels (blind set for testing),** which is used to evaluate the submitted systems. The total number of 10-s segments is 7920 (22 hours), which is significantly larger than the Development dataset.

The Framework Proposed – *Experimental Setting*

TABLE II
THE NUMBER OF 10-SECOND AUDIO RECORDINGS CORRESPONDING TO EACH SCENE CATEGORIES IN THE TRAIN. AND EVAL. SUBSETS SEPARATED FROM THE DCASE 2021 TASK 1A DEVELOPMENT DATASET [35], AND THE EVALUATION DATASET [32].

Category	Train. Subset	Eval. Subset	Evaluation
Airport	1393	296	-
Bus	1400	297	-
Metro	1382	297	-
Metro Station	1380	297	-
Park	1429	297	-
Public square	1427	297	-
Shopping mall	1373	297	-
Street pedestrian	1386	297	-
Street traffic	1413	297	-
Tram	1379	296	-
Total	13962 (≈38.79 hours)	2968 (≈8.25 hours)	7920 (22 hours)

Acoustic Scene Classification (ASC),
Unbalanced Dataset,
Mismatched Recording Devices

* The **Development dataset (Train. & Eval. Subsets):** The dataset in slightly unbalanced, recorded from 12 large European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna.
The audio recordings were recorded from 3 different devices namely A (10215 recordings), B (749 recordings), C (748 recordings). Additionally, synthetic data for 11 mobile devices was created based on the original recordings, referred to as S1 (750 recordings), S2 (750 recordings), S3 (750 recordings), S4 (750 recordings), S5 (750 recordings), and S6 (750 recordings).

* The **Evaluation dataset without labels (blind set for testing),** which is used to evaluate the submitted systems. The total number of 10-s segments is 7920 (22 hours), which is significantly larger than the Development dataset.

The Framework Proposed – *Experimental Setting*

TABLE II
THE NUMBER OF 10-SECOND AUDIO RECORDINGS CORRESPONDING TO EACH SCENE CATEGORIES IN THE TRAIN. AND EVAL. SUBSETS SEPARATED FROM THE DCASE 2021 TASK 1A DEVELOPMENT DATASET [35], AND THE EVALUATION DATASET [32].

Category	Train. Subset	Eval. Subset	Evaluation
Airport	1393	296	-
Bus	1400	297	-
Metro	1382	297	-
Metro Station	1380	297	-
Park	1429	297	-
Public square	1427	297	-
Shopping mall	1373	297	-
Street pedestrian	1386	297	-
Street traffic	1413	297	-
Tram	1379	296	-
Total	13962 (≈38.79 hours)	2968 (≈8.25 hours)	7920 (22 hours)

Acoustic Scene Classification (ASC),
Unbalanced Dataset,
Mismatched Recording Devices

* The **Development dataset (Train. & Eval. Subsets):** The dataset in slightly unbalanced, recorded from 12 large European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna.
The audio recordings were recorded from 3 different devices namely A (10215 recordings), B (749 recordings), C (748 recordings). Additionally, synthetic data for 11 mobile devices was created based on the original recordings, referred to as S1 (750 recordings), S2 (750 recordings), S3 (750 recordings), S4 (750 recordings), S5 (750 recordings), and S6 (750 recordings).

* The **Evaluation dataset without labels (blind set for testing),** which is used to evaluate the submitted systems. The total number of 10-s segments is 7920 (22 hours), which is significantly larger than the Development dataset.

* Librosa for generating spectrogram,
Tensorflow framework for constructing Neural Networks,
Kullback-Leibler (KL) loss function, Adam optimization,
ACC (%) metric, ….

Agenda

1. Motivation & Dataset

2. The Framework Proposed

3. Results & Discussion

Results

Performance comparison of CNN-7 w/ CR & DC among three spectrograms, with different time lengths, with or without using data augmentations (Acc. %)

Results

Performance comparison of CNN-7 w/ CR & DC among three spectrograms, with different time lengths, with or without using data augmentations (Acc. %)

	Without data augmentations				With Data augmentations			
Spectrogram	**1-second**	**2-second**	**5-second**	**10-second**	**1-second**	**2-second**	**5-second**	**10-second**
MEL	56.7	57.9	56.2	60.5	54.6	57.9	59.5	58.4
GAM	53.2	55.0	53.1	53.9	58.9	60.1	60.6	57.1
CQT	44.3	47.7	48.6	49.2	44.2	45.7	48.6	49.1

Results

Performance comparison of CNN-7 w/ CR & DC among three spectrograms, with different time lengths, with or without using data augmentations (Acc. %)

	Without data augmentations				With Data augmentations			
Spectrogram	**1-second**	**2-second**	**5-second**	**10-second**	**1-second**	**2-second**	**5-second**	**10-second**
MEL	56.7	57.9	56.2	60.5	54.6	57.9	59.5	58.4
GAM	53.2	55.0	53.1	53.9	58.9	60.1	60.6	57.1
CQT	44.3	47.7	48.6	49.2	44.2	45.7	48.6	49.1

Results

Performance comparison of CNN-7 w/ CR & DC among three spectrograms, with different time lengths, with or without using data augmentations (Acc. %)

	Without data augmentations				With Data augmentations			
Spectrogram	**1-second**	**2-second**	**5-second**	**10-second**	**1-second**	**2-second**	**5-second**	**10-second**
MEL	56.7	57.9	56.2	60.5	54.6	57.9	59.5	58.4
GAM	53.2	55.0	53.1	53.9	58.9	60.1	60.6	57.1
CQT	44.3	47.7	48.6	49.2	44.2	45.7	48.6	49.1

Performance comparison among DCASE baseline, the CNN-7 baseline, the cnn-7 baseline with channel restriction (CNN-7 w/ CR), the cnn-7 baseline with channel restriction and decomposed convolution (CNN-7 w/ CR & DC) (Acc. %).

Results

Performance comparison of CNN-7 w/ CR & DC among three spectrograms, with different time lengths, with or without using data augmentations (Acc. %)

	Without data augmentations				With Data augmentations			
Spectrogram	**1-second**	**2-second**	**5-second**	**10-second**	**1-second**	**2-second**	**5-second**	**10-second**
MEL	56.7	57.9	56.2	60.5	54.6	57.9	59.5	58.4
GAM	53.2	55.0	53.1	53.9	58.9	60.1	60.6	57.1
CQT	44.3	47.7	48.6	49.2	44.2	45.7	48.6	49.1

Performance comparison among DCASE baseline, the CNN-7 baseline, the cnn-7 baseline with channel restriction (CNN-7 w/ CR), the cnn-7 baseline with channel restriction and decomposed convolution (CNN-7 w/ CR & DC) (Acc. %).

Category	**DCASE baseline (90.3 KB)**	**CNN-7 baseline (1.1 MB)**	**CNN-7 w/ CR (313 KB)**	**CNN-7 w/ CR & DC (42.6 KB)**
Airport	40.5	59.5	50.3	64.5
Bus	47.1	73.7	70.4	69.0
Metro	51.9	57.6	49.8	70.0
Metro station	28.3	53.9	48.1	45.1
Park	69.0	73.1	78.5	74.4
Public square	25.3	34.3	38.4	25.9
Shopping mall	61.3	52.9	50.2	43.4
Street pedestrian	38.7	39.4	35.0	32.7
Street traffic	62.0	84.5	88.2	89.6
Tram	53.0	67.9	62.5	52.7
Average	47.7	59.7	57.1	56.7

Results

Performance comparison of CNN-7 w/ CR & DC among three spectrograms, with different time lengths, with or without using data augmentations (Acc. %)

	Without data augmentations				With Data augmentations			
Spectrogram	**1-second**	**2-second**	**5-second**	**10-second**	**1-second**	**2-second**	**5-second**	**10-second**
MEL	56.7	57.9	56.2	60.5	54.6	57.9	59.5	58.4
GAM	53.2	55.0	53.1	53.9	58.9	60.1	60.6	57.1
CQT	44.3	47.7	48.6	49.2	44.2	45.7	48.6	49.1

Performance comparison among DCASE baseline, the CNN-7 baseline, the cnn-7 baseline with channel restriction (CNN-7 w/ CR), the cnn-7 baseline with channel restriction and decomposed convolution (CNN-7 w/ CR & DC) (Acc. %).

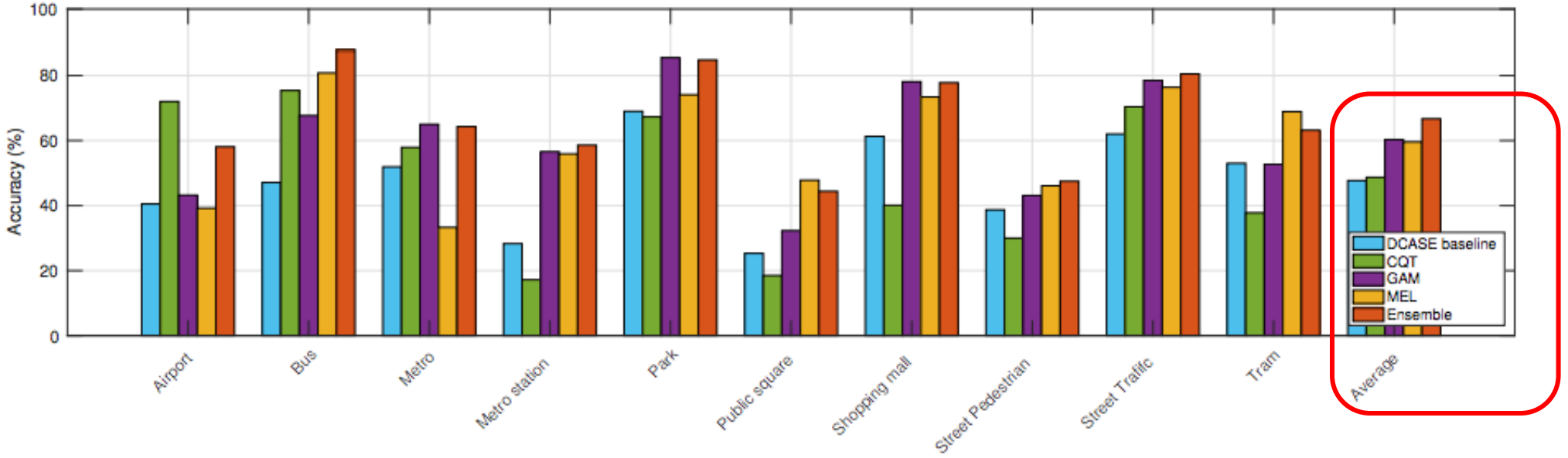
Category	**DCASE baseline (90.3 KB)**	**CNN-7 baseline (1.1 MB)**	**CNN-7 w/ CR (313 KB)**	**CNN-7 w/ CR & DC (42.6 KB)**
Airport	40.5	59.5	50.3	64.5
Bus	47.1	73.7	70.4	69.0
Metro	51.9	57.6	49.8	70.0
Metro station	28.3	53.9	48.1	45.1
Park	69.0	73.1	78.5	74.4
Public square	25.3	34.3	38.4	25.9
Shopping mall	61.3	52.9	50.2	43.4
Street pedestrian	38.7	39.4	35.0	32.7
Street traffic	62.0	84.5	88.2	89.6
Tram	53.0	67.9	62.5	52.7
Average	47.7	59.7	57.1	56.7

Results

Performance comparison (Acc.%) of DCASE baseline, individual spectrograms (CQT, GAM, and MEL), and the ensemble of three spectrograms across all scene categories (using CNN-7 with CR & DC, 5-second time length, and mixup & spectrum data augmentations)

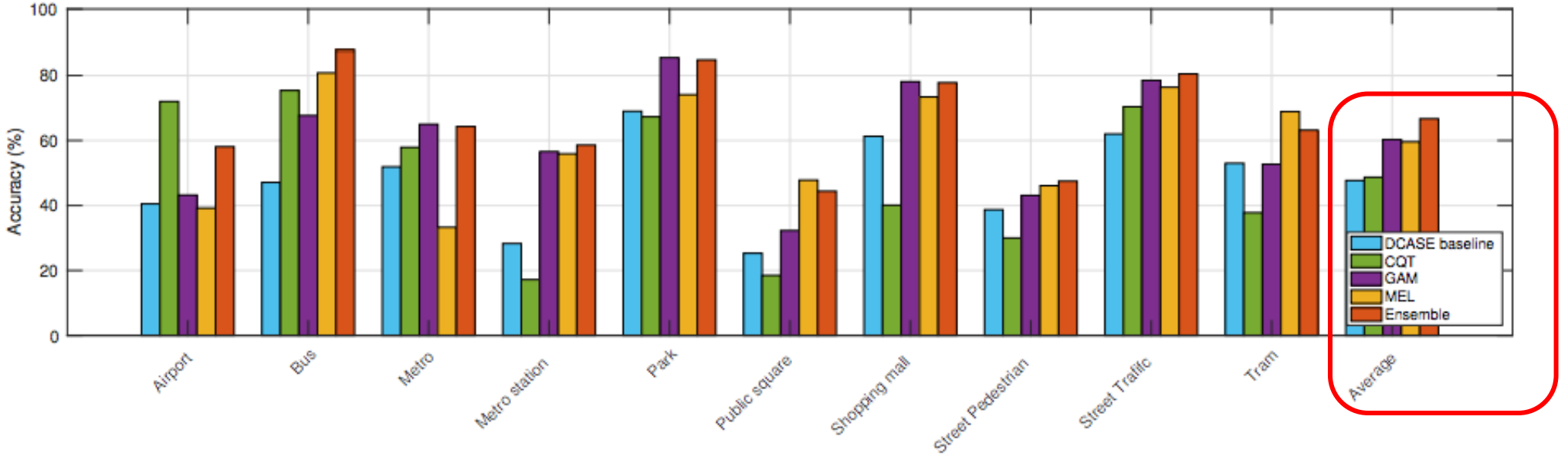
Results

Performance comparison (Acc.%) of DCASE baseline, individual spectrograms (CQT, GAM, and MEL), and the ensemble of three spectrograms across all scene categories (using CNN-7 with CR & DC, 5-second time length, and mixup & spectrum data augmentations)



Results

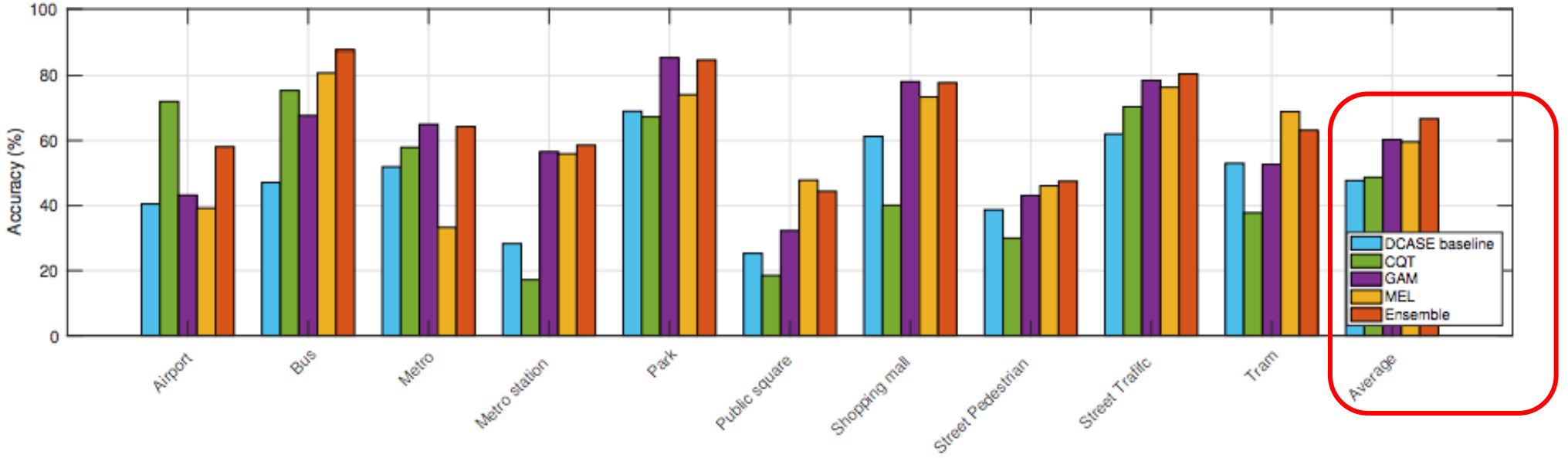
Performance comparison (Acc.%) of DCASE baseline, individual spectrograms (CQT, GAM, and MEL), and the ensemble of three spectrograms across all scene categories (using CNN-7 with CR & DC, 5-second time length, and mixup & spectrum data augmentations)



The number of 10-second audio scene recordings corresponding to each device in the train. and eval. subsets separated from the DCASE 2021 Task 1a development dataset [35] and performance for each devices.

Results

Performance comparison (Acc.%) of DCASE baseline, individual spectrograms (CQT, GAM, and MEL), and the ensemble of three spectrograms across all scene categories (using CNN-7 with CR & DC, 5-second time length, and mixup & spectrum data augmentations)

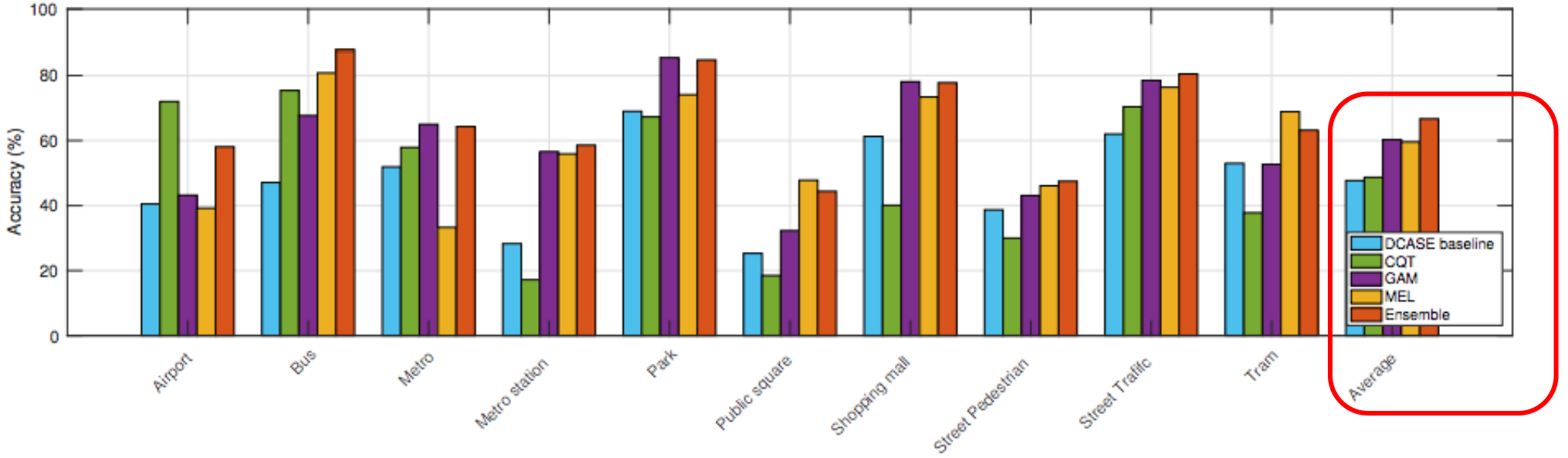


The number of 10-second audio scene recordings corresponding to each device in the train. and eval. subsets separated from the DCASE 2021 Task 1a development dataset [35] and performance for each devices.

Devices	Train.	Eval.	Acc. %
A	10215	330	79.1
B	749	329	69.6
C	748	329	70.8
S1	750	330	65.8
S2	750	330	63.6
S3	750	330	67.0
S4	0	330	63.9
S5	0	330	60.0
S6	0	330	60.3

Results

Performance comparison (Acc.%) of DCASE baseline, individual spectrograms (CQT, GAM, and MEL), and the ensemble of three spectrograms across all scene categories (using CNN-7 with CR & DC, 5-second time length, and mixup & spectrum data augmentations)



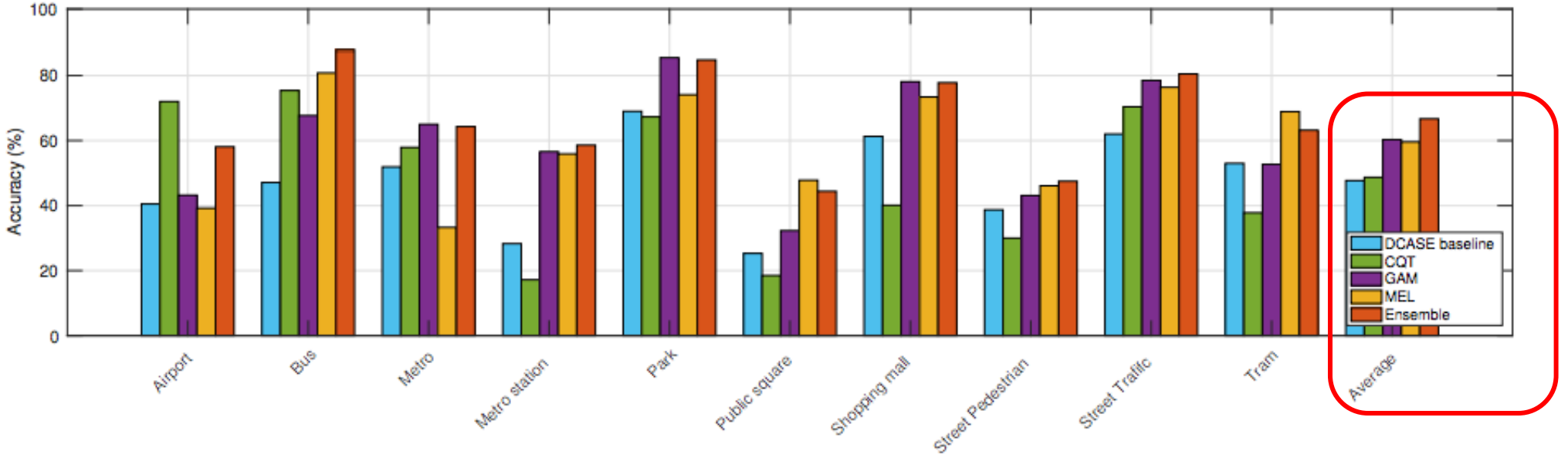
The number of 10-second audio scene recordings corresponding to each device in the train. and eval. subsets separated from the DCASE 2021 Task 1a development dataset [35] and performance for each devices.

Devices	Train.	Eval.	Acc. %
A	10215	330	79.1
B	749	329	69.6
C	748	329	70.8
S1	750	330	65.8
S2	750	330	63.6
S3	750	330	67.0
S4	0	330	63.9
S5	0	330	60.0
S6	0	330	60.3

Top-10 accuracy performance (Acc. %) systems submitted for DCASE 2021 Task 1a challenge

Results

Performance comparison (Acc.%) of DCASE baseline, individual spectrograms (CQT, GAM, and MEL), and the ensemble of three spectrograms across all scene categories (using CNN-7 with CR & DC, 5-second time length, and mixup & spectrum data augmentations)



The number of 10-second audio scene recordings corresponding to each device in the train. and eval. subsets separated from the DCASE 2021 Task 1a development dataset [35] and performance for each devices.

Devices	Train.	Eval.	Acc. %
A	10215	330	79.1
B	749	329	69.6
C	748	329	70.8
S1	750	330	65.8
S2	750	330	63.6
S3	750	330	67.0
S4	0	330	63.9
S5	0	330	60.0
S6	0	330	60.3

Top-10 accuracy performance (Acc. %) systems submitted for DCASE 2021 Task 1a challenge

Systems	Evaluation dataset	Eva. Subset
Top-1 [40]	76.1	75.9
Top-2 [41]	72.9	-
Top-3 [42]	72.1	69.5
Top-4 [43]	70.3	69.0
Top-5 [44]	70.1	-
Our system	**69.6**	**66.7**
Top-7 [45]	69.6	65.0
Top-8 [46]	68.8	70.2
Top-9 [47]	68.5	65.2
Top-10 [48]	68.3	69.7
DCASE baseline [31]	45.6	47.7

Discussion

* Achievements:

Multiple spectrograms, data augmentation, model compression to deal with ASC challenges for high-performance low-complexity model (i.e. target to edge devices)

Discussion

* Achievements:

Multiple spectrograms, data augmentation, model compression to deal with ASC challenges for high-performance low-complexity model (i.e. target to edge devices)

*Further research on:

+ Techniques of model compression: Distillation, Pruning, Quantization, etc.
+ Novel neural network architectures to improve performance.
+ Issue of mismatched recording devices.

Thank you & Questions