Unsupervised Contrastive Learning of Sound Event Representations

Eduardo Fonseca[1]*, Diego Ortego[2]*, Kevin McGuinness[2], Noel E. O'Connor[2] and Xavier Serra[1]

*Equal contribution - Paper ID: 4255

https://github.com/edufonseca/uclser20

Yerun Young European Research Universities

1 Universitat Pompeu Fabra Barcelona

MTG Music Technology Group

2 Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS

ICASSP 2021 TORONTO Canada June 6-11, 2021 Metro Toronto Convention Centre

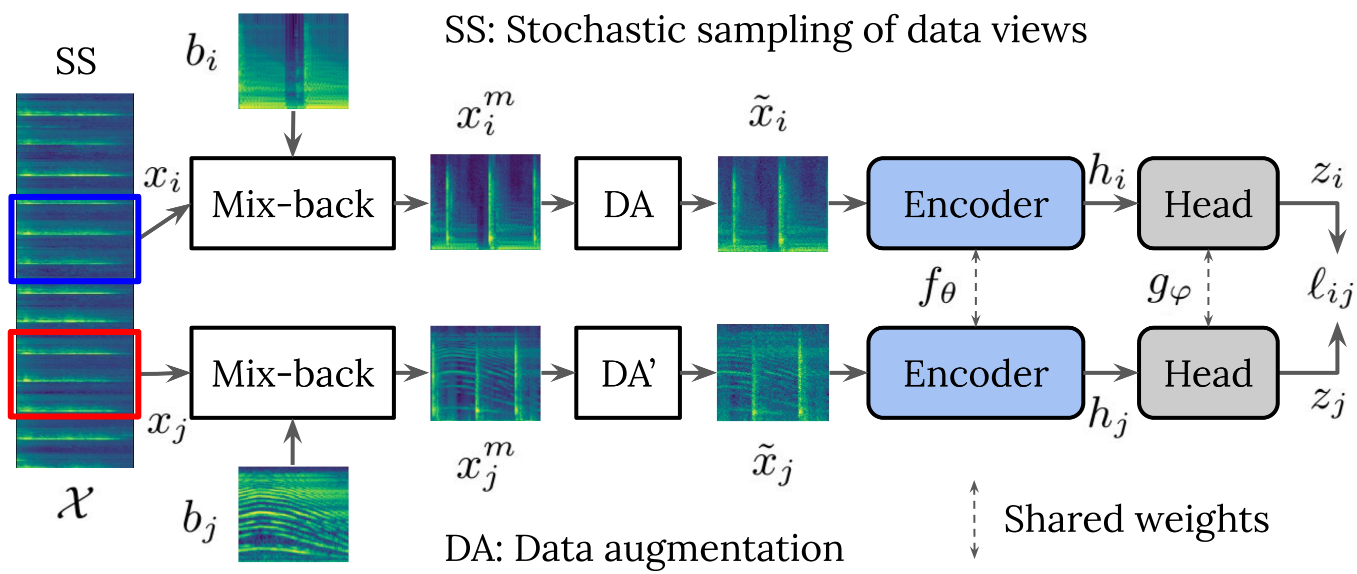
Context

- **Task**: learn sound event representation in unsupervised fashion
- **Motivation**: common scenario in sound event research
  - → few manually labeled data but **abundant unlabeled data**
- Self-supervised learning
  - → Learn representation from unlabeled data without explicit labels
  - → Generate pseudo-labels, $\hat{y}$, from the data itself
  - → Key factor: design **proxy task** to generate $\hat{y}$ → useful representations emerge

Contrastive Representation Learning

- Contrastive learning is learning by comparing
  - We compare between pairs of input examples:
    - **positive** pairs of **similar** inputs
    - **negative** pairs of **unrelated** inputs
- Goal is an embedding space where representations …
  - of **similar** examples ➜ **close** together
  - of **dissimilar** examples ➜ **further** away

Proposed Approach: Overview



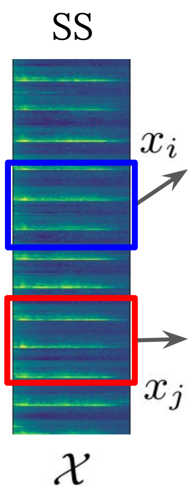
Proxy task

- Similarity maximization, inspired by SimCLR [1]
  - maximize similarity between differently augmented views of sound events
- Input: log-mel spectrograms
- Output: embedding representations *h*

[1] Chen et al., **A Simple Framework for Contrastive Learning of Visual Representations**. ICML 2020

Proposed Approach: Sampling TF patches

SS: Stochastic sampling of data views

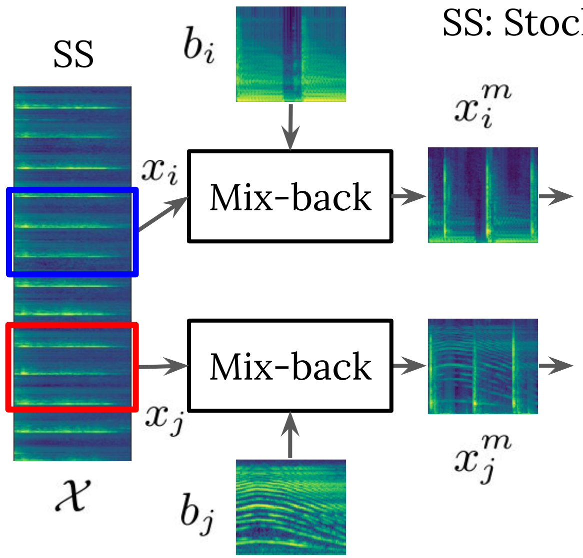


Sampling TF patches (aka Temporal Proximity [2])

- Sample two patches (views) at random within audio clip log-mel spectrogram
- TxF=101x96
- Temporal coherence among neighbouring patches ➔ natural data augmentation
  - same source / different pattern
  - different source related semantically

[2] Jansen et al., **Unsupervised learning of semantic audio representations.** ICASSP 2018

Proposed Approach: *mix-back*



SS: Stochastic sampling of data views

$$x_i^m = (1 - \lambda)\, x_i + \lambda \left[E(x_i) / E(b_i) \right] b_i$$

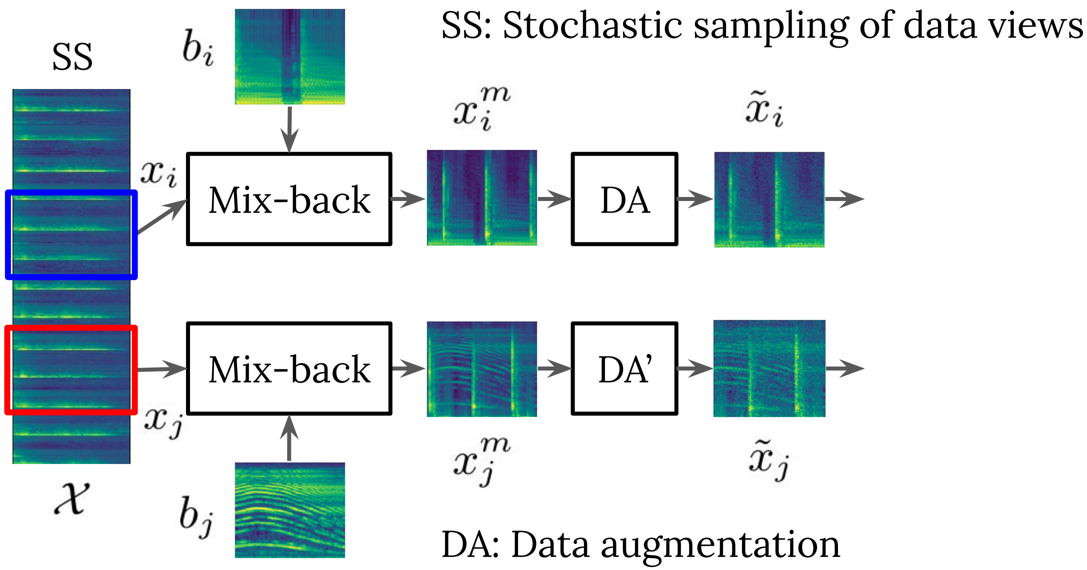
$$\lambda \sim \mathcal{U}(0, \alpha) \quad \alpha \in [0, 1]$$

$E(\cdot)$ is energy

Mix **incoming patch with a *back*ground patch**

- Goal:
  - reduce mutual information via mixing with random backgrounds
  - keeping relevant semantics by sound transparency
- Energy (E) adjustment ensures that x_i is always dominant over b_i
- Prevent aggressive transformations that may make the proxy task too difficult

Proposed Approach: Data Augmentation

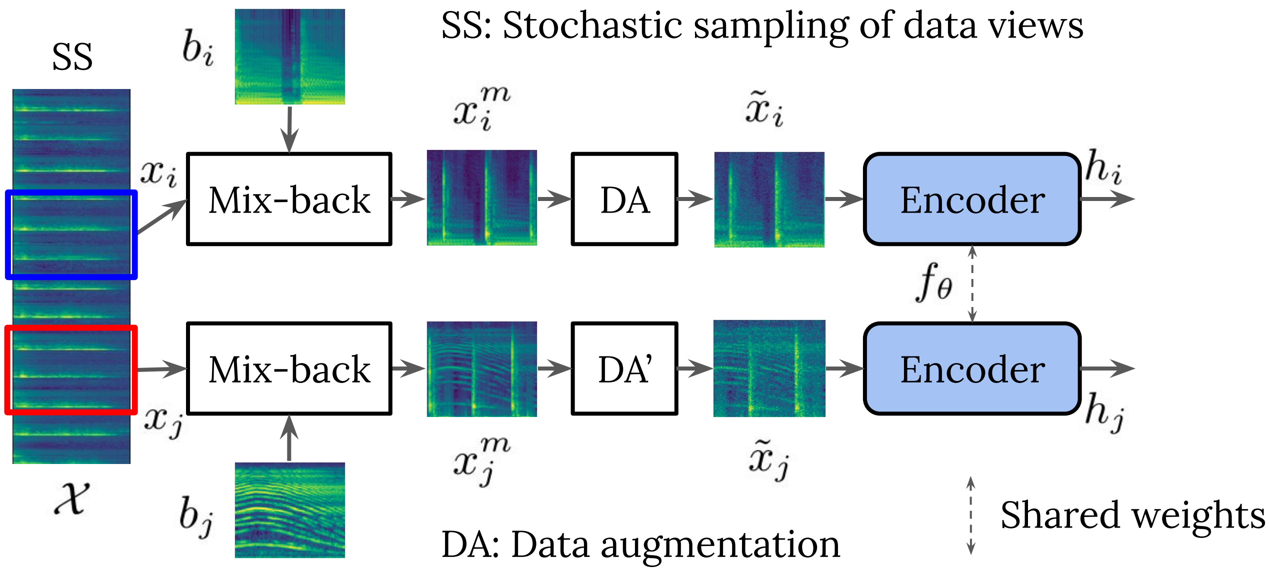


Stochastic Data Augmentation

- Directly over TF patches
- Simple for on-the-fly computation
- **Random resized cropping (RRC)**, **compression**, **Gaussian noise addition**, **specAugment [3]**, random time/frequency shifts, Gaussian blurring
- Hyper-parameters randomly sampled from a distribution for each patch

[3] Park et al., **SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition.** InterSpeech 2019

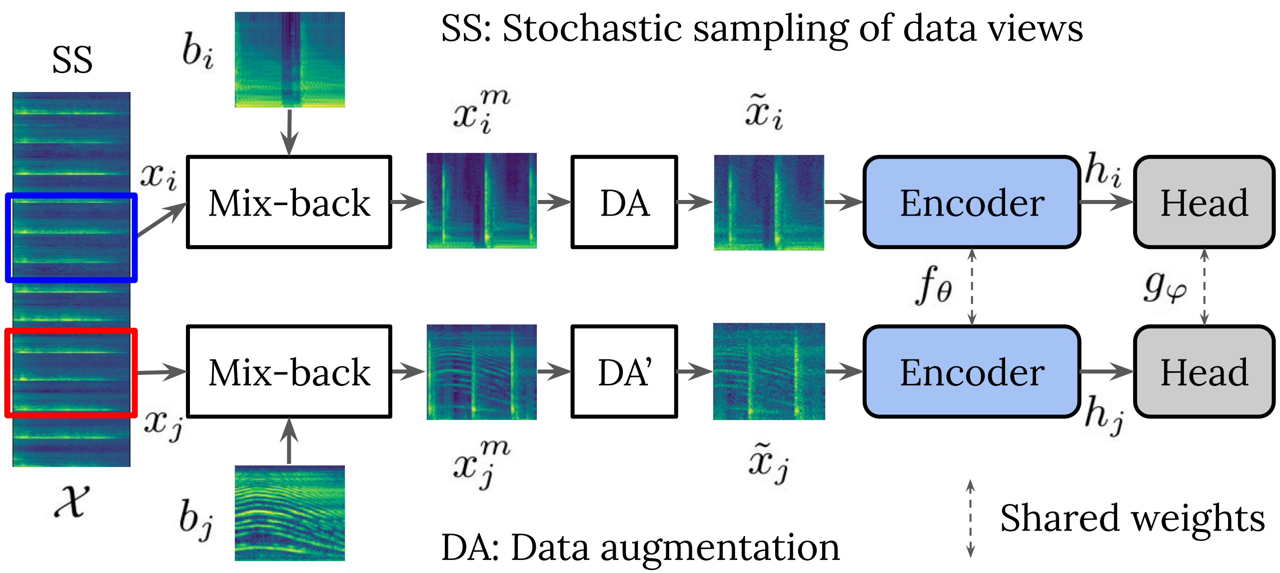
Proposed Approach: Encoder



Convolutional encoder

- Extract low-dimensional embeddings *h*
- Once the training is over, *h* is used for downstream tasks
- ResNet-18 / VGG-like / CRNN after removing classification layer

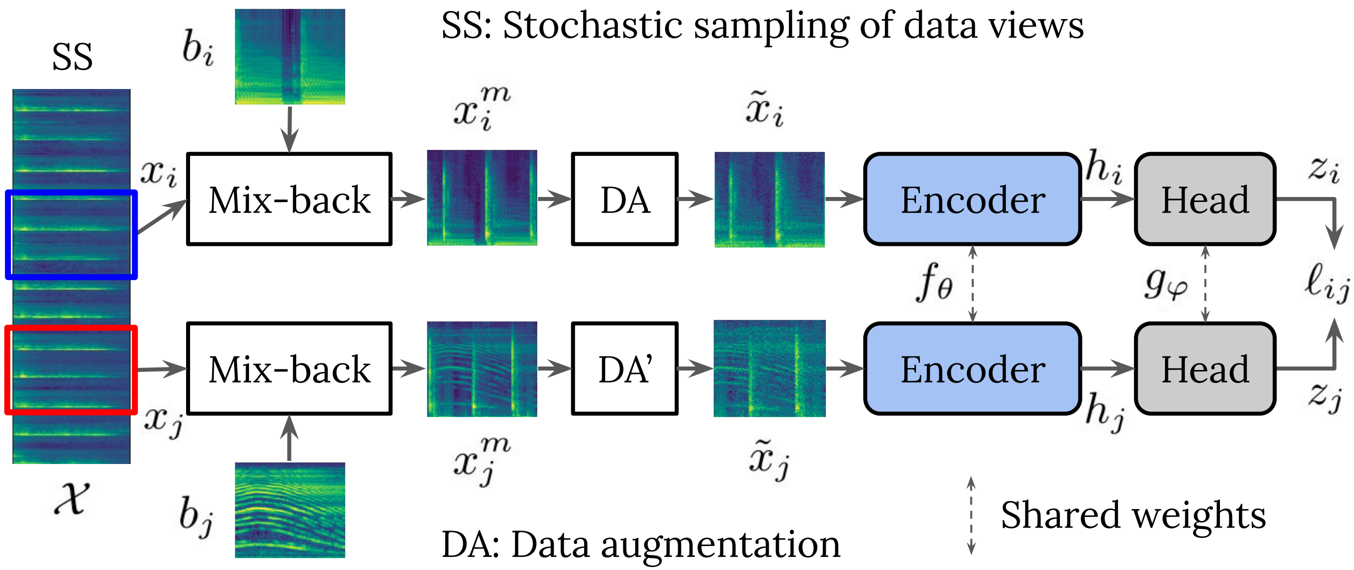
Proposed Approach: Head



Projection Head

- Map *h* to L2-normalized metric embedding *z*, where loss is applied
- MLP w/ one hidden layer + BNorm + ReLU

Proposed Approach: Contrastive Loss



Normalized temperature-scaled cross-entropy (NT-Xent) loss [1]

- Softmax structure
- Scoring function: cosine similarity with temperature scaling $\boldsymbol{\tau}$
- Maximize similarity between differently augmented views

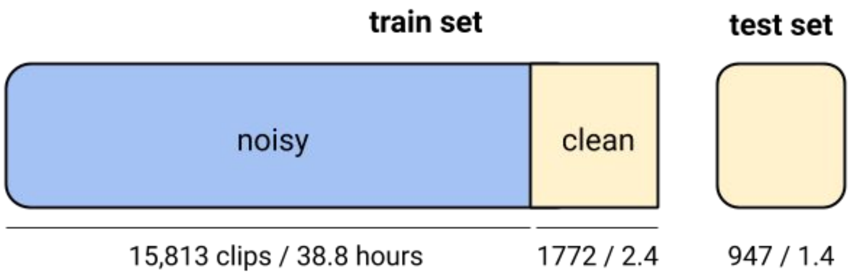
$$\ell_{ij} = -\log \frac{\exp\left(sim(z_i, z_j)/\tau\right)}{\sum_{v=1}^{2N} \mathbb{1}_{v \neq i} \exp\left(sim(z_i, z_v)/\tau\right)}$$

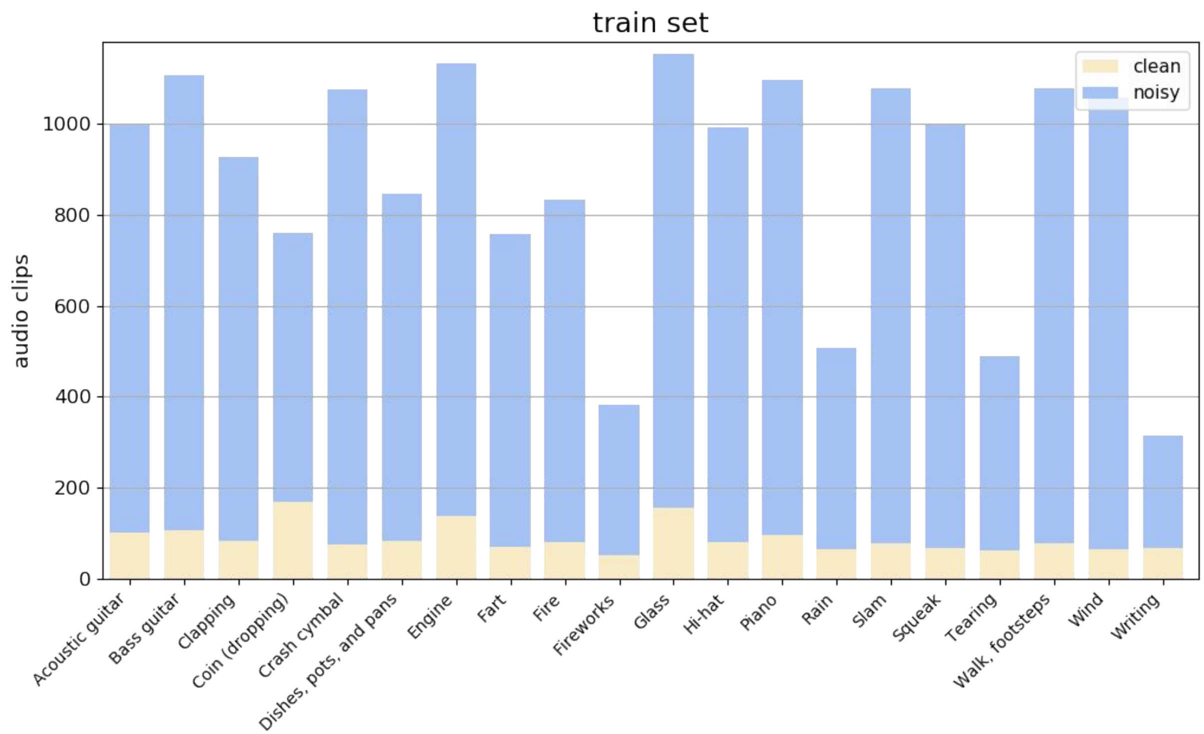
[1] Chen et al., **A Simple Framework for Contrastive Learning of Visual Representations**. ICML 2020

Evaluation: FSDnoisy18k dataset

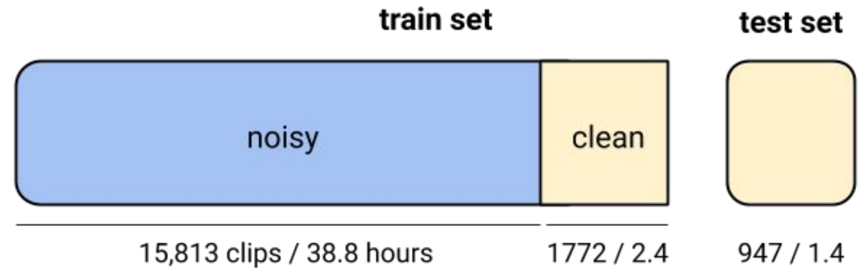
www.eduardofonseca.net/FSDnoisy18k/

- 20 classes / 18k clips / 42.5 h [4]
- singly-labeled data ➔ accuracy as metric
- proportion train_noisy / train_clean = 90% / 10%
- per-class varying degree of label noise





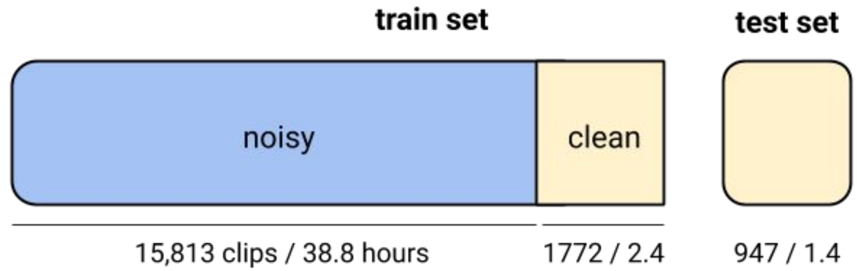
Evaluation Methodology



Two stages

1. **Unsupervised representation learning**
   - train on *train_noisy* without labels
   - validate on *train_clean* using labels in **kNN Evaluation**:
     - estimate representation *z* for each patch
     - pairwise cosine similarity with rest of patches
     - prediction by majority voting across k=200 neighbouring labels

Evaluation Methodology



Two stages

1. **Unsupervised representation learning**
   - train on *train_noisy* without labels
   - validate on *train_clean* using labels in **kNN Evaluation**:
     - estimate representation *z* for each patch
     - pairwise cosine similarity with rest of patches
     - prediction by majority voting across k=200 neighbouring labels

2. **Evaluation of the representation** using **supervised** tasks (w/ labels)
   - **Linear Evaluation**: train **additional linear classifier** on top of pre-trained unsupervised embeddings
     - train on train_noisy / validate on train_clean
   - **End-to-end Fine Tuning**: **fine-tune model** on two downstream tasks after initializing with pre-trained weights:
     1. train on train_noisy / validate on train_clean
     2. train on train_clean (allow 15% for validation)

Ablation Study: Sampling TF patches

- **best**: sampling at random
- **worst**: using same patch
- overlapping patches (d < 101 frames) ➜ detrimental
- results accord with [5]
- effective method used in most contrastive learning approaches for audio representation learning

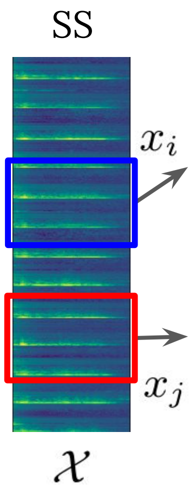


Table 1. kNN val accuracy for several ways of sampling TF patches.

Sampling method	kNN	Sampling method	kNN
Sampling at random	**70.1**	$d = 125$	67.9
$d = 0$ (same patch)	51.1	$d = 200$	69.9
$d = 25$	61.5	$d = 300$	68.5
$d = 75$	65.1	$d = 400$	69.7

[5] Tian et al., **What Makes for Good Views for Contrastive Learning?** NeurIPS 2020

Ablation Study: mix-back

- lightly mixing patches with real backgrounds from unrelated patches helps
- adjusting the energy is also beneficial
  - foreground patch is dominant over the background patch
  - preventing aggressive transforms & keeping semantics

Table 2. kNN val accuracy for several mix-back and data augmentation (DA) settings.

Mix-back setting (α)	**kNN**
w/ E adjustment (0.05)	**70.1**
w/o E adjustment (0.02)	66.2
w/o mix-back	63.3

Ablation Study: Data Augmentation (DA)

→ Each row: best result after sweeping the corresponding parameters

1. Explore DAs applied **individually**
   - random resized cropping: small stretch in time/freq & small freq transposition
   - SpecAugment (time/freq masking) [3]

Table 2. kNN val accuracy for several mix-back and data augmentation (DA) settings.

DA policy	kNN
RRC + comp + noise	**70.1**
RRC + comp	69.6
RRC + specAugment	70.0
RRC	69.0
specAugment [20]	68.0
w/o DA	60.1

[3] Park et al., **SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition**. InterSpeech 2019

Ablation Study: Data Augmentation (DA)

→ Each row: best result after sweeping the corresponding parameters

1. Explore DAs applied **individually**
   - random resized cropping: small stretch in time/freq & small freq transposition
   - SpecAugment (time/freq masking) [3]
2. Explore DA **compositions** based on RRC
   - RRC + compression + Gaussian noise addition
   - RRC + SpecAugment
   - more exhaustive exploration of the DA compositions ➔ better results

Table 2. kNN val accuracy for several mix-back and data augmentation (DA) settings.

DA policy	kNN
RRC + comp + noise	**70.1**
RRC + comp	69.6
RRC + specAugment	70.0
RRC	69.0
specAugment [20]	68.0
w/o DA	60.1

[3] Park et al., **SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition.** InterSpeech 2019

Evaluation of Learned Representations

Supervised baselines & Linear Evaluation

→ Supervised baselines: CRNN ≊ VGG-like > ResNet-18
  - ResNet-18: large capacity for not so much data & noisy labels

Model	Linear	Supervised baseline
(weights in M)	-	
ResNet-18 (11)	**74.3**	65.4
VGG-like (0.3)	70.0	70.6
CRNN (1)	64.4	72.0

Evaluation of Learned Representations

Supervised baselines & Linear Evaluation

- Supervised baselines: CRNN ≊ VGG-like > ResNet-18
  - ResNet-18: large capacity for not so much data & noisy labels
- Linear Evaluation:
  - ResNet-18 is top
    - larger capacity is better for unsupervised contrastive learning
    - exceeds supervised performance
  - VGG-like & CRNN: most of the supervised performance is recovered

Model	**Linear**	**Supervised baseline**
(weights in M)	-	
ResNet-18 (11)	**74.3**	65.4
VGG-like (0.3)	70.0	70.6
CRNN (1)	64.4	72.0

Evaluation of Learned Representations

Fine tuning on downstream tasks after initializing with pre-trained weights

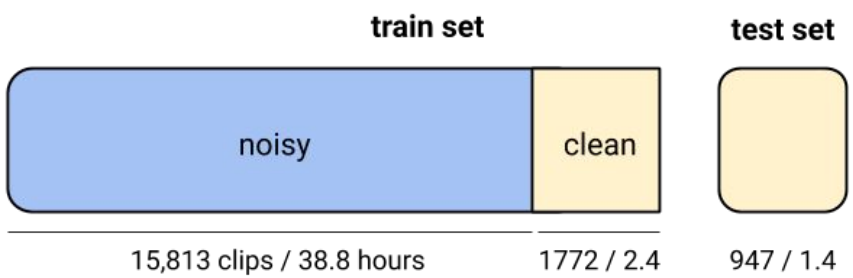
- Goal: measure benefit wrt training from scratch in noisy- & small-data regimes
- **Unsupervised contrastive pre-training is best in all cases**
- ResNet-18:
  - **lowest** accuracy trained from **scratch** (limited by data or label quality)
  - **top** accuracy w/ unsupervised **pre-training** (alleviate these problems)
- Greater improvements in "smaller clean" task

Model	**Linear**	**Larger noisy set**		**Small clean set**	
(weights in M)		random*	p-t	random	p-t
ResNet-18 (11)		65.4	**78.2**	56.5	**77.9**
VGG-like (0.3)		70.6	**72.8**	61.1	**72.3**
CRNN (1)		72.0	**74.2**	58.7	**69.1**

Evaluation of Learned Representations

Fine tuning on downstream tasks after initializing with pre-trained weights

- Pre-trained performance ➔ little degradation between tasks: why?
  - “smaller clean” task: fine tune on **unseen clean** data (albeit small)
  - “larger noisy” task: fine tune on **same** data used for unsupervised learning (now affected by **label noise**)



Model	**Linear**	**Larger noisy set**		**Small clean set**	
(weights in M)		random*	p-t	random	p-t
ResNet-18 (11)		65.4	**78.2**	56.5	**77.9**
VGG-like (0.3)		70.6	**72.8**	61.1	**72.3**
CRNN (1)		72.0	**74.2**	58.7	**69.1**

Summary & Takeaways

- Framework for unsupervised contrastive learning of sound event representations
- Maximize similarity between differently augmented views of the same spectrogram
- Successful representation learning by tuning compound
  - positive patch sampling & mix-back & data augmentation
- Unsupervised contrastive pre-training can
  - mitigate the impact of data scarcity
  - increase robustness against noisy labels
- Fine tuning a model initialized with pretrained weights outperforms supervised baselines

Unsupervised Contrastive Learning of Sound Event Representations

Thank you!

Q/A?

Eduardo Fonseca[1]*, Diego Ortego[2]*, Kevin McGuinness[2],
Noel E. O'Connor[2] and Xavier Serra[1]

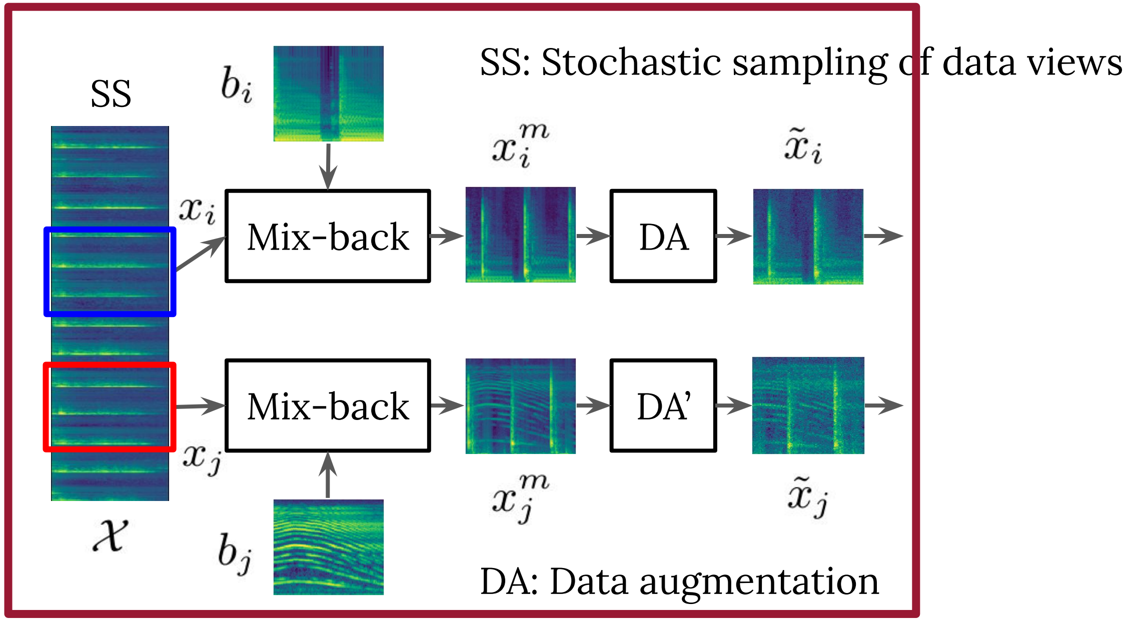
*Equal contribution - Paper ID: 4255

https://github.com/edufonseca/uclser20

Yerun Young European Research Universities
1 upf. Universitat Pompeu Fabra Barcelona
MTG Music Technology Group
2 Insight SFI RESEARCH CENTRE FOR DATA ANALYTICS

ICASSP 2021 TORONTO Canada June 6-11, 2021
Metro Toronto Convention Centre

Proposed Approach: Data Augmentation



Generating views for contrastive learning of audio representations

1. Sampling patches
2. mix-back
3. Basic augmentations

Ablation Study: Discussion

- Framework is **sensitive**
  - compositions, parameter tuning, $\boldsymbol{\tau}$, etc
  - not one key ingredient but a compound

Ablation Study: Discussion

- Framework is **sensitive**
  - compositions, parameter tuning, τ , etc
  - not one key ingredient but a compound
- **Composing augmentations helps**, but done carefully
  - ordering of the DAs matter
  - joining individually-tuned DAs can be suboptimal (affect each other)
  - tuning composition can be computationally intensive

Ablation Study: Discussion

- Framework is **sensitive**
  - compositions, parameter tuning, τ , etc
  - not one key ingredient but a compound
- **Composing augmentations helps**, but done carefully
  - ordering of the DAs matter
  - joining individually-tuned DAs can be suboptimal (affect each other)
  - tuning composition can be computationally intensive
- Hypothesis: ***shortcuts*** mitigated by sampling patches and mix-back
  - time-frequency patterns used to lower the loss w/o useful learning
    - recording gear, room acoustics, background, …

Ablation Study: Discussion

- Framework is **sensitive**
  - compositions, parameter tuning, τ , etc
  - not one key ingredient but a compound
- **Composing augmentations helps**, but done carefully
  - ordering of the DAs matter
  - joining individually-tuned DAs can be suboptimal (affect each other)
  - tuning composition can be computationally intensive
- Hypothesis: ***shortcuts*** mitigated by sampling patches and mix-back
  - time-frequency patterns used to lower the loss w/o useful learning
    - recording gear, room acoustics, background, ...
- **Batch size**:
  - common knowledge: the larger the better (more negative examples)
  - our case: batch size of 128 (worse scenario)