IEEE ICASSP 2021• 6-11 June 2021 • Toronto, Ontario, Canada

Automatic multitrack mixing with a differentiable mixing console of neural audio effects

Christian J. Steinmetz[1,2] Jordi Pons[1] Santiago Pascual[1] Joan Serrà[1]

[1]Dolby Laboratories
[2]Music Technology Group, Universitat Pompeu Fabra, Barcelona

What is (automatic) mixing?

Plugin processors
(compressor, EQ, reverb, etc.)

Stereo panning
(imaging and spatialisation)

Level faders
(relative volume control)

Expert systems
(Knowledge engineering)

vs.

Machine Learning
(Classical ML algorithms)

Pro: Produces explainable decisions
Con: Lacks sufficient complexity

(De Man and Reiss, 2013)

Pro: Provides greater model flexibility
Con: Complete absence of parametric data

(Moffat and Sandler, 2019)

These systems fail to generalize to real-world music production

Automatic mixing references: https://csteinmetz1.github.io/AutomaticMixingPapers/

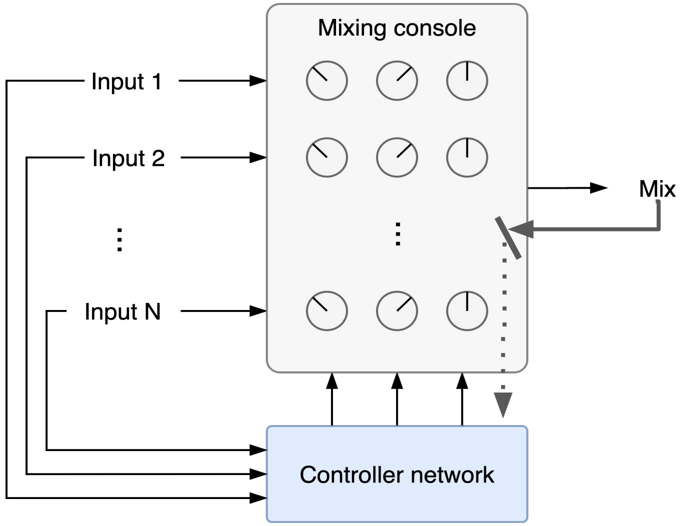
Can deep learning enable us to learn mixing techniques directly from tracks and mixes without the underlying mixing parameters?

Key challenges

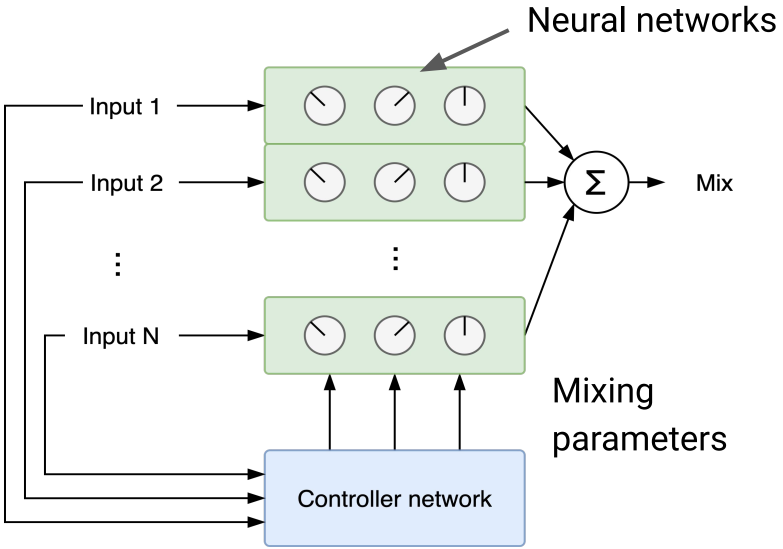
In the application of deep learning for mixing

1. Evaluation of mixes — What makes a good mix? According to who?
2. Highly variable inputs — No consistent size and structure to inputs.
3. High-fidelity required — High sampling rates and no artifacts.
4. User interaction — Audio engineers need to tweak the output.

We could use traditional DSP effects as a strong inductive bias for the mixing task

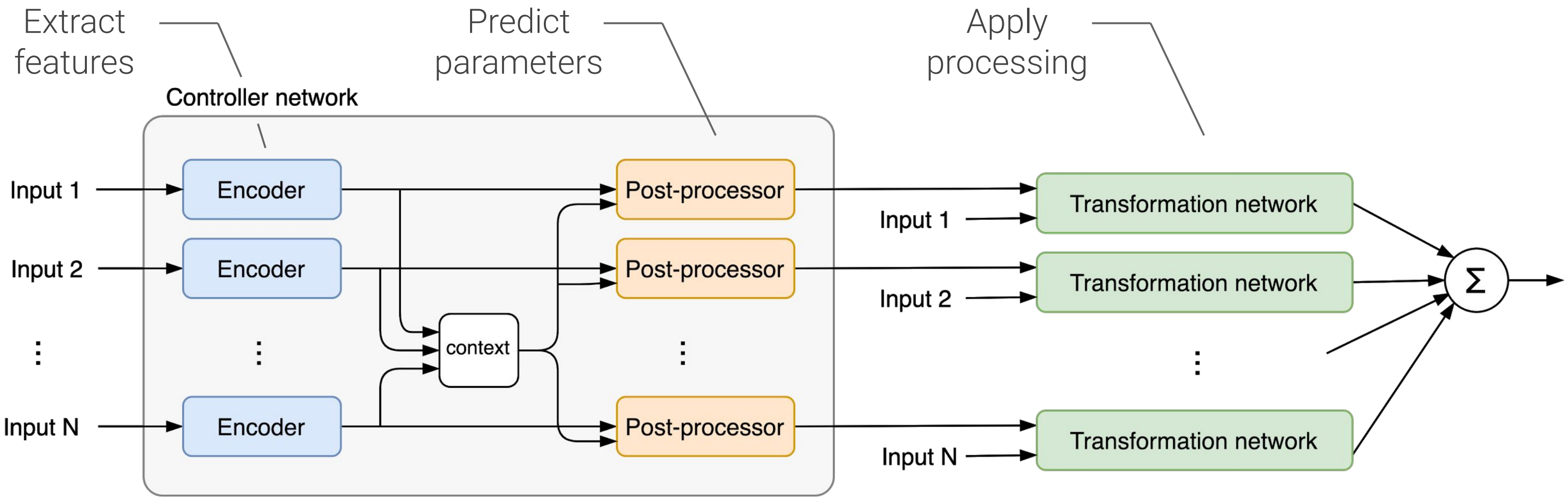


Unfortunately, the mixing console is not differentiable



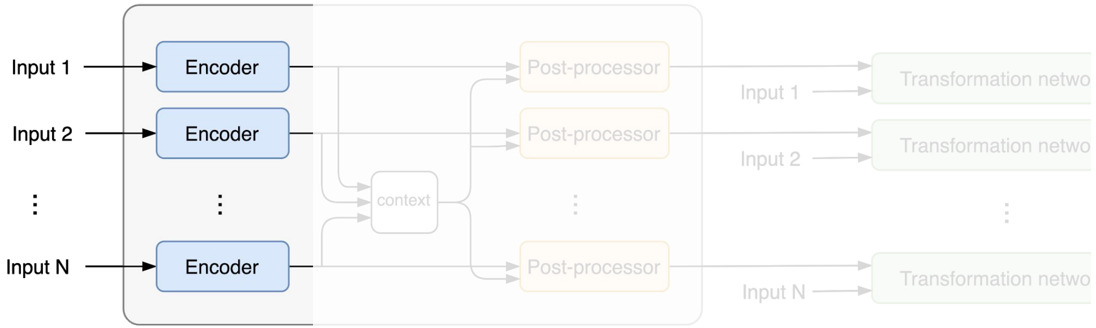
...but we can train a differentiable model to emulate a channel.

Differentiable mixing console

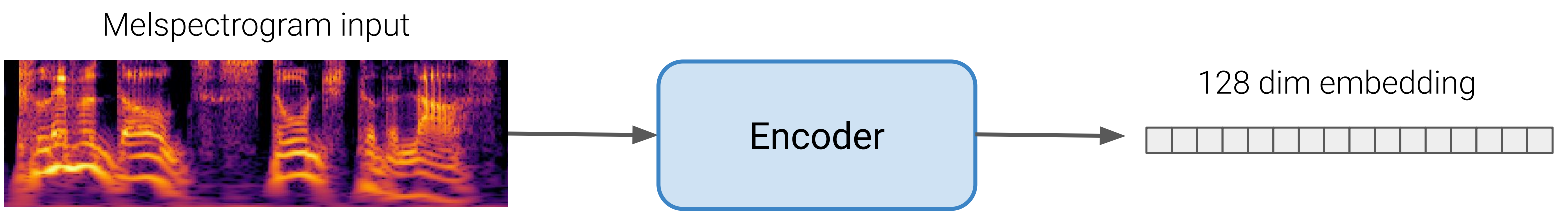


Encoder

Extract info from inputs for making mixing decisions



Generates 128 dim embedding for each input channel

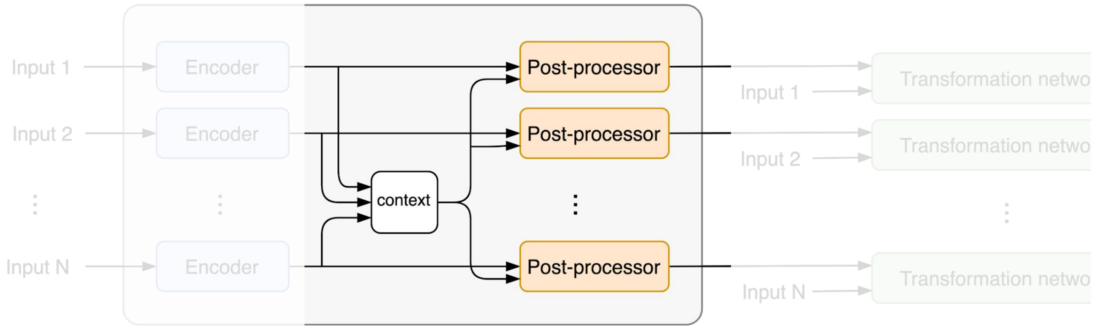


VGGish trained on AudioSet

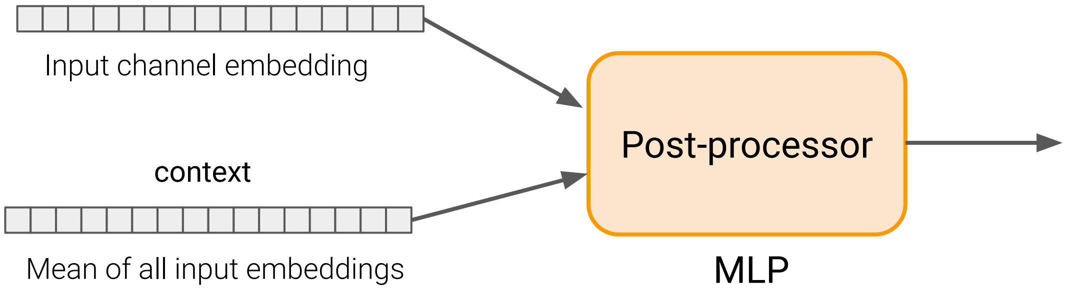
(Hershey et al., 2017)

Post-processor

Aggregate information to make mixing decisions



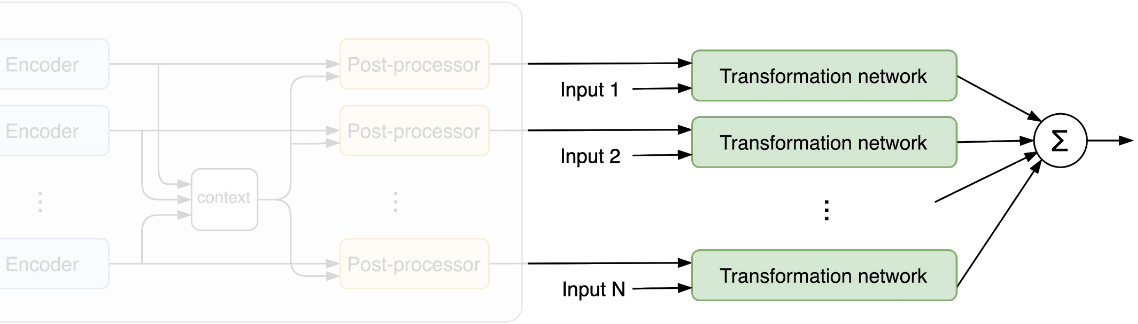
A single MLP is distributed across all input channels (shared weights). This provides input ordering invariance and places no limit on number of input channels.



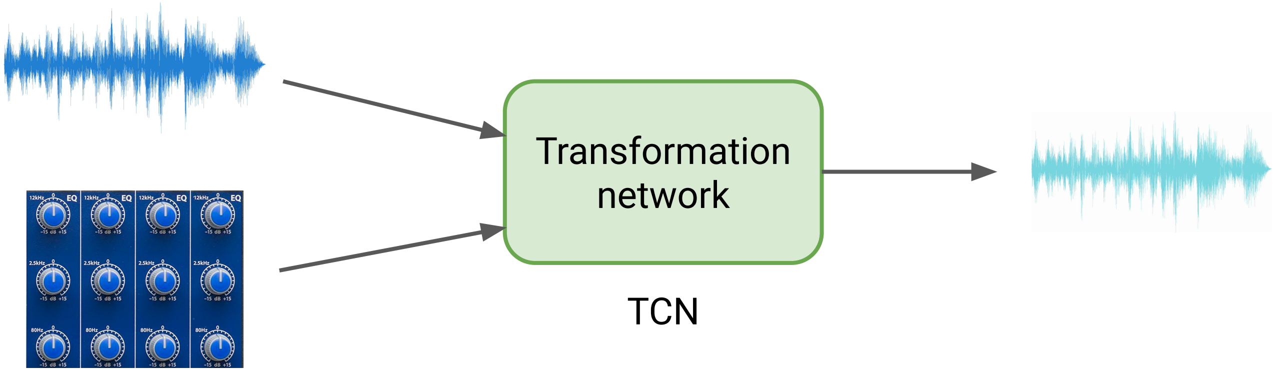
Mix parameters
for a single channel

Transformation network

Perform the types of processing employed in mixing
(but in a differentiable framework)



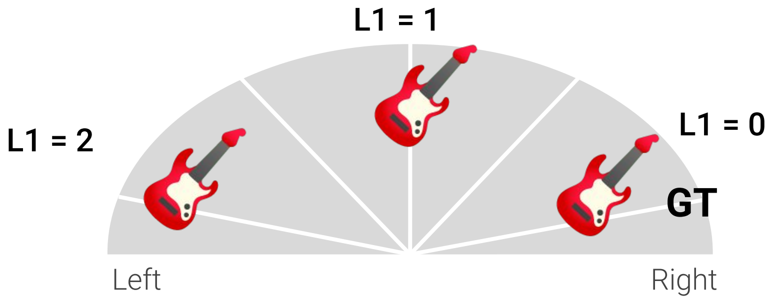
We can train a model on this model with generated data by processing audio files with random settings of existing audio effects.



Stereo loss function

Loss function to encourage realistic mixes

Panning here is more perceptually similar but gives a higher L1 loss



L1 and L2 loss on stereo signals encourage panning all elements to the center.

$$y_{\text{sum}} = y_{\text{left}} + y_{\text{right}}$$

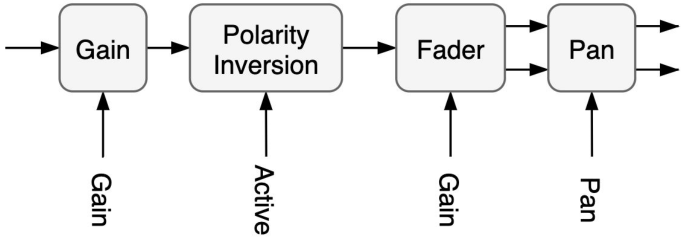
$$y_{\text{diff}} = y_{\text{left}} - y_{\text{right}}$$

$$\ell_{\text{Stereo}}(\hat{y}, y) = \ell_{\text{MR-STFT}}(\hat{y}_{\text{sum}}, y_{\text{sum}}) + \ell_{\text{MR-STFT}}(\hat{y}_{\text{diff}}, y_{\text{diff}})$$

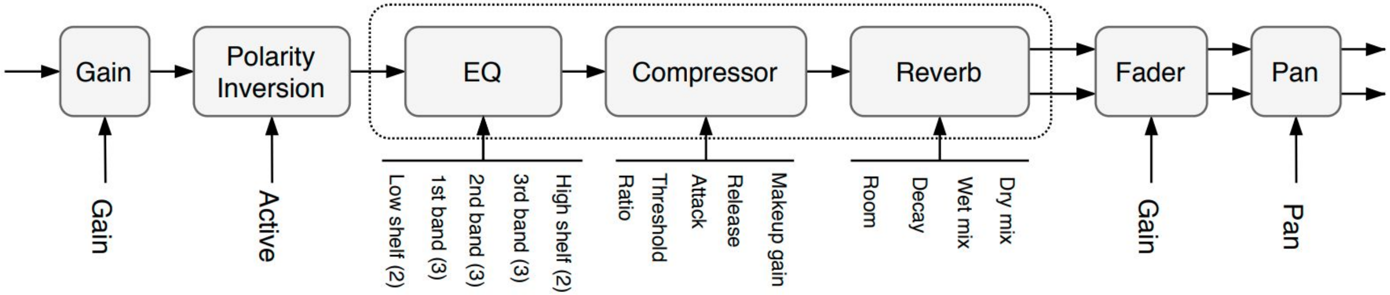
Achieves invariance to stereo (left-right) orientation

Model configurations

Gain + Panning (Transformation network is not used)



Gain + EQ + Compressor + Reverb + Panning



Datasets

ENST-drums

(Gillet and Richard, 2006)

Easier, but less realistic mixing task

Recordings from three drummers, all follow same 8 channel structure

MedleyDB

(Bittner et al., 2016)

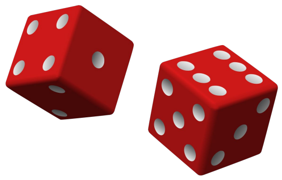
Challenging, but realistic mixing task

Diverse styles, varying number of tracks (2-100), complete songs

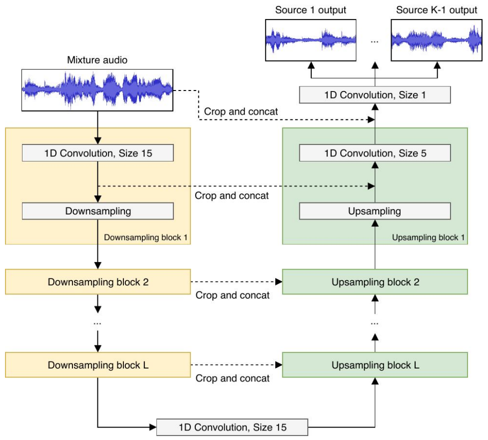
Baselines

Mono mix

Random mix



Wave-U-Net mix



(Stoller et al., 2018)

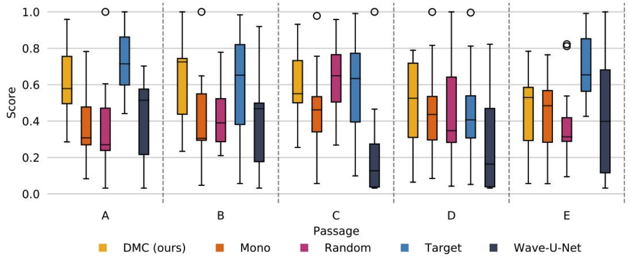
(Martínez Ramírez et al., 2021)

Demo

Perceptual evaluation

ENST-drums (8 channels)

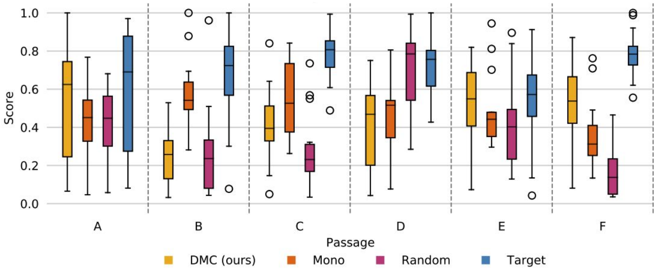
Gain + panning configuration



Listeners rate mixes from our system higher than baselines in the drum mixing task.

MedleyDB (6 channels)

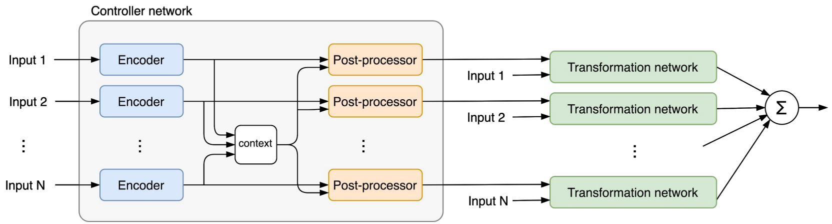
Gain + EQ + Compressor + Reverb + Panning



Our mixes often exceed baselines, but creating mixes with all the processors is a lot harder...

Contributions

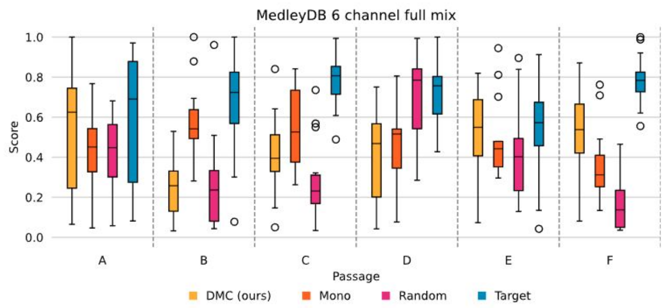
Deep learning based multitrack mixing system



Our end-to-end mixing architecture:

1. Can be trained with a limited number of examples
2. Learns mixing conventions directly from stereo mixes
3. Makes no assumptions about input sources
4. Places no limit on the number of input sources
5. Enables users to adjust the mix results (interpretability)

MedleyDB full mixing task



Passage	DMC (ours)	Mono	Random	Target
A	0:00 / 0:29	0:00 / 0:29	0:00 / 0:29	0:00 / 0:29
B	0:00 / 0:14	0:00 / 0:14	0:00 / 0:14	0:00 / 0:14
C	0:00 / 0:29	0:00 / 0:29	0:00 / 0:29	0:00 / 0:29
D	0:00 / 0:29	0:00 / 0:29	0:00 / 0:29	0:00 / 0:29
E	0:00 / 0:28	0:00 / 0:28	0:00 / 0:28	0:00 / 0:28
F	0:00 / 0:14	0:00 / 0:14	0:00 / 0:14	0:00 / 0:14

See the companion website for more listening examples

https://csteinmetz1.github.io/dmc-icassp2021

IEEE ICASSP 2021• 6-11 June 2021 • Toronto, Ontario, Canada

Automatic multitrack mixing with a differentiable mixing console of neural audio effects

Christian J. Steinmetz[1,2] Jordi Pons[1] Santiago Pascual[1] Joan Serrà[1]

[1]Dolby Laboratories
[2]Music Technology Group, Universitat Pompeu Fabra, Barcelona