The Inverse Drum Machine: Source Separation Through Joint Transcription and Analysis-by-Synthesis

Bernardo Torres Geoffroy Peeters Gaël Richard

LTCI, Télécom Paris, Institut Polytechnique de Paris

Audio demos & code

Key Contributions

- **Multitask Optimization**:
  - Automatic Drum Transcription (ADT)
  - One-shot drum Sample Synthesis (OSS)
  - Drum Source Separation (DSS)
- **No Isolated Stems Required**: an end-to-end, **Analysis-by-Synthesis** separation model for drums trained on drum mixtures and transcription annotations.
- **High Efficiency**: separation quality comparable to supervised SOTA with **100x fewer parameters** (465K vs 49.1M).

1 Drum Source Separation

- **The challenge:** isolated stem recordings for drums are heavy, inefficient and almost impossible to obtain.

Synthesis inductive bias ⇒ no need for isolated stems

- **The drum machine model**: discrete one-shot samples triggered at onset times.
- We **learn** to **invert** this process by training a one-shot synthesis model and a transcription model end-to-end.

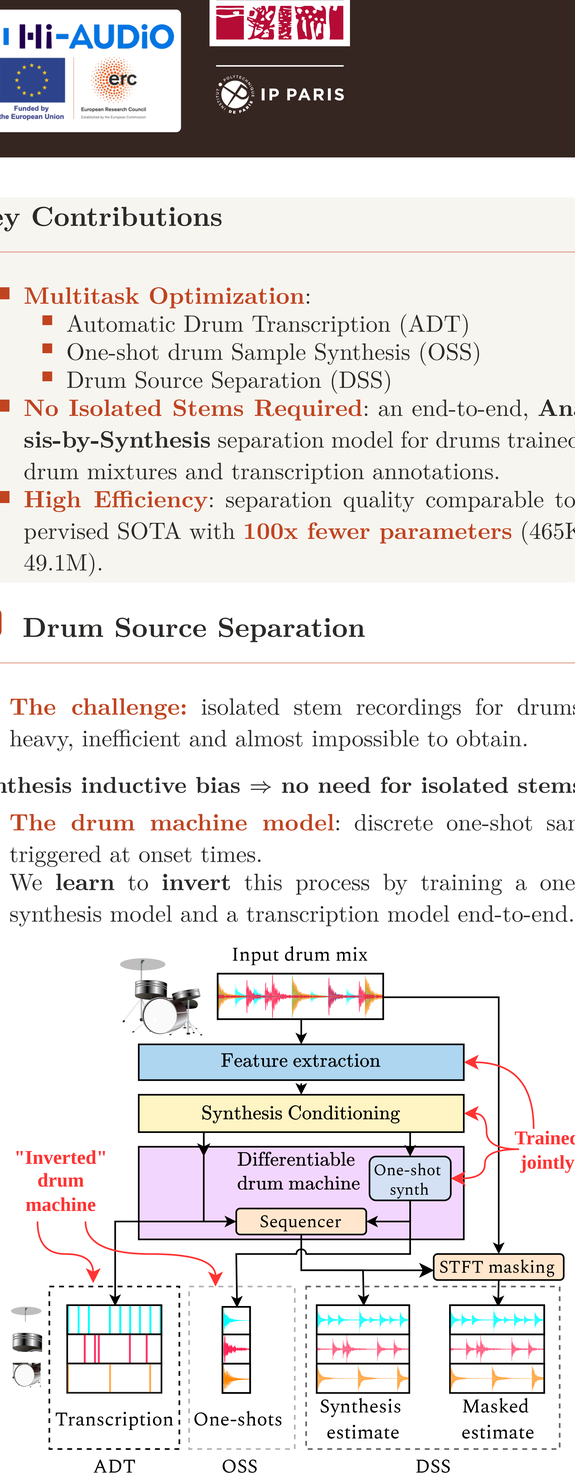


Figure 1: The modular network architecture of the Inverse Drum Machine (IDM). Outputs compose three tasks: Automatic Drum Transcription (ADT), One-shot Sample Synthesis (OSS), and Drum Source Separation (DSS).

- **Masking (optional)**: α-Wiener masking at inference

2 Drum One-shot Synthesis

A causal Temporal Convolutional Network (TCN) conditioned on a mixture embedding and instrument class.

- Noise → conditioned TCN → exponential envelope → 1-second one-shots
- Convolved with onsets in differentiable sequencer to reconstruct drum track. It **is never shown ground-truth one-shots** as target.

3 Analysis-by-Synthesis, multitask learning

- **Synthesis Parameters**: onsets, velocities, mixture embedding, and per-track gains

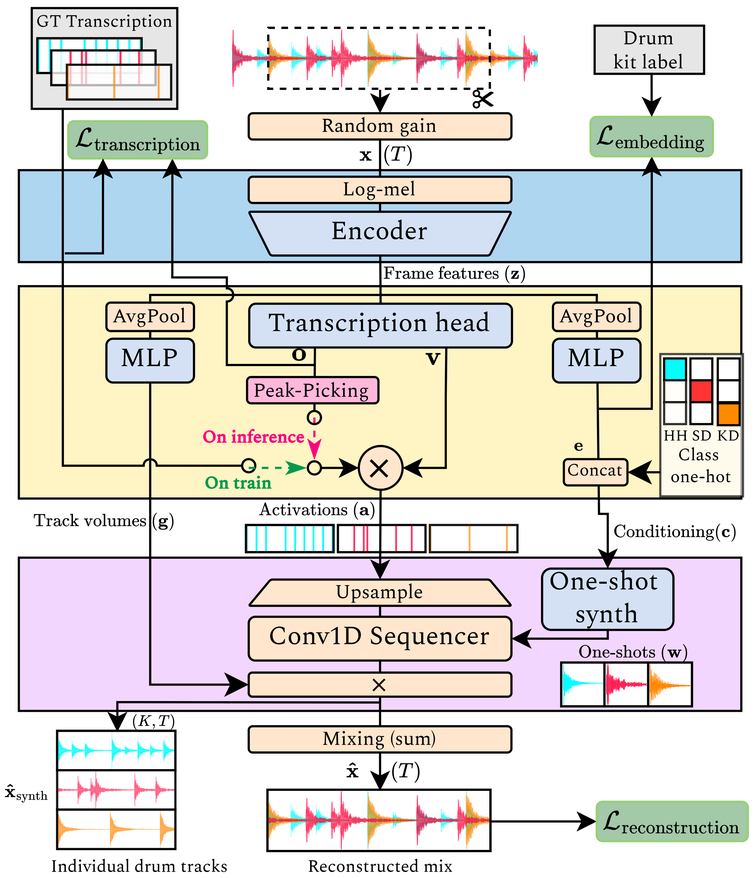
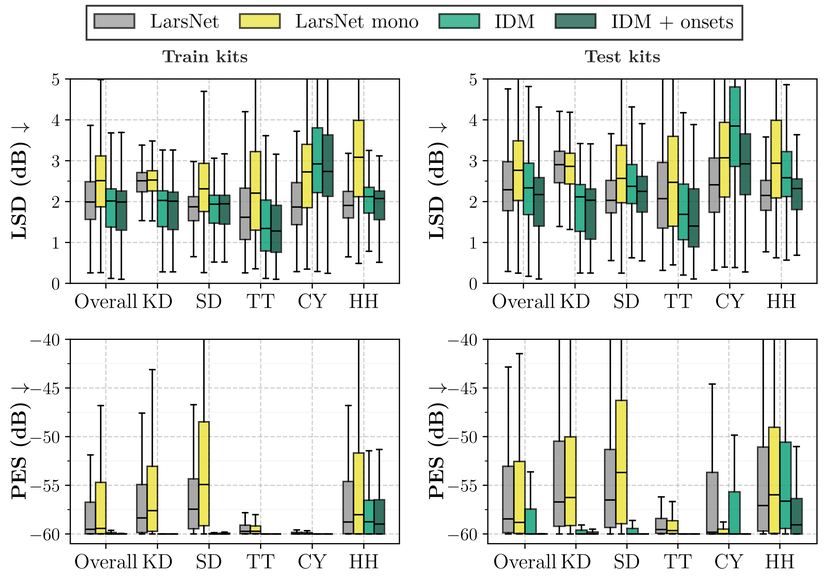


Figure 2: Blue = trainable; orange = differentiable DSP; pink = non-differentiable; green = losses.

4 Separation Results on Direct Synthesis



* LSD = Log Spectral Distance, PES = Predicted Energy on Silence. Lower is better for both.

- Better spectral similarity and silence prediction on many classes.
- Even though IDM can't generate new one-shots, separation results on unseen kits are still strong.

5 Separation Results after Wiener Masking

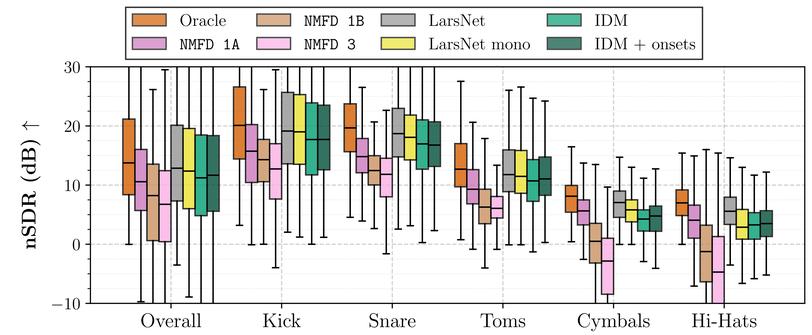


Figure 3: nSDR per instrument class. IDM (ours, green) achieves comparable quality to the supervised LarsNet and better performance than NMFD methods with transcription information.

- Comparable nSDR and Log Spectral Distance (LSD) to supervised SOTA (LarsNet - Mezza et al, 2024) with 100× fewer parameters (465K vs 49.1M) and only transcription supervision.
- **Modular architecture allows us to override onsets at inference time** (+ onsets) which improves results. Transcription quality is a key bottleneck.