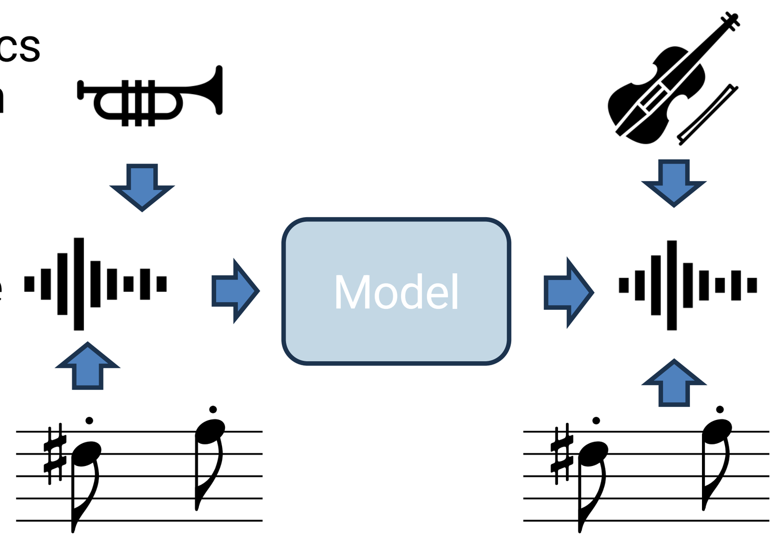
TIMBRE TRANSFER USING IMAGE-TO-IMAGE DENOISING DIFFUSION IMPLICIT MODELS

Luca Comanducci, Fabio Antonacci, Augusto Sarti
Politecnico di Milano

ISMIR 2023 Milan, Italy
Nov. 5-9, 2023

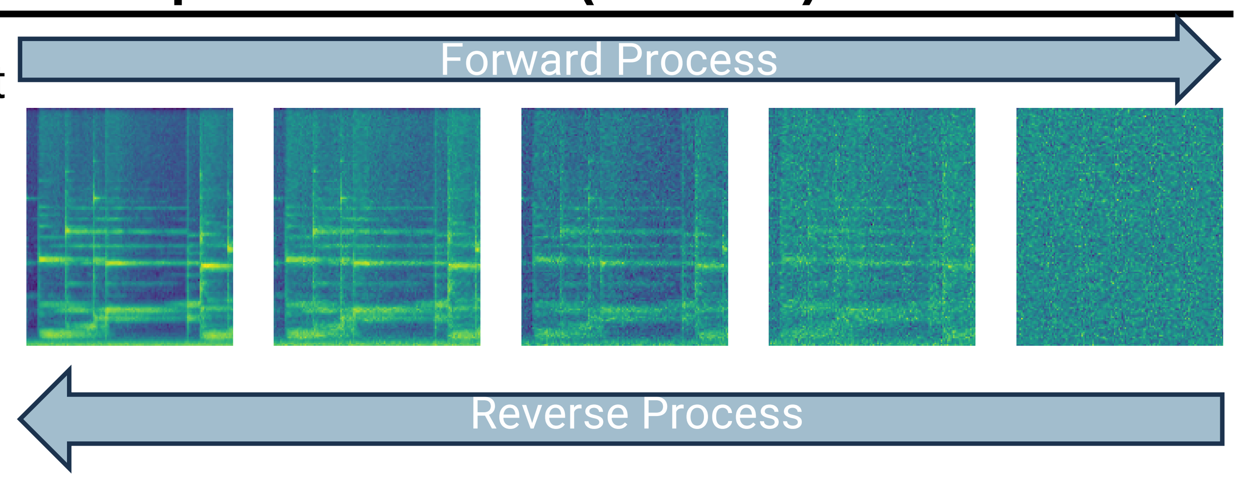
Context

- **Musical Timbre** is the ""perceived characteristics of a musical sound that are different from pitch and amplitude contours"" [1].
- **Timbre Transfer** consists in converting a musical piece from one timbre to another while preserving the other music-related characteristics.
- Usually performed through generative models such as Generative Adversarial Networks (CycleGAN)
- In this work we apply Denoising Diffusion Models



Denoising Diffusion Implicit Models (DDIMs)

- **Diffusion Models** convert input samples from a standard Gaussian distribution into samples from an empirical data distribution through iterative denoising process
  - Forward Process → adding noise
  - Reverse Process → Removing noise (U-Net)

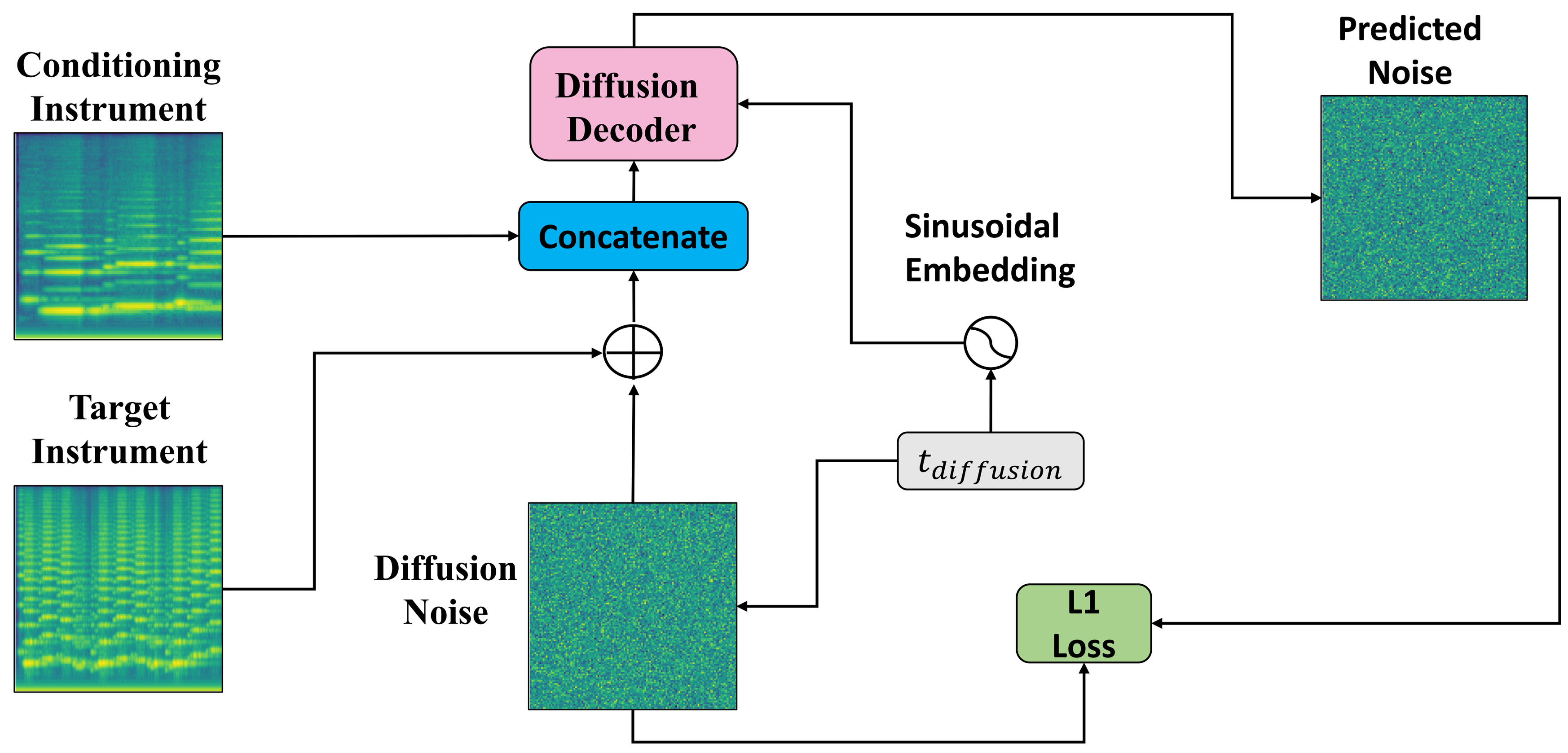


- **Denoising Diffusion Implicit Models [2]**
  - Generalize to non-markovian forward diffusion process
  - Same training procedure of probabilistic counterpart
  - Allow for faster sampling times

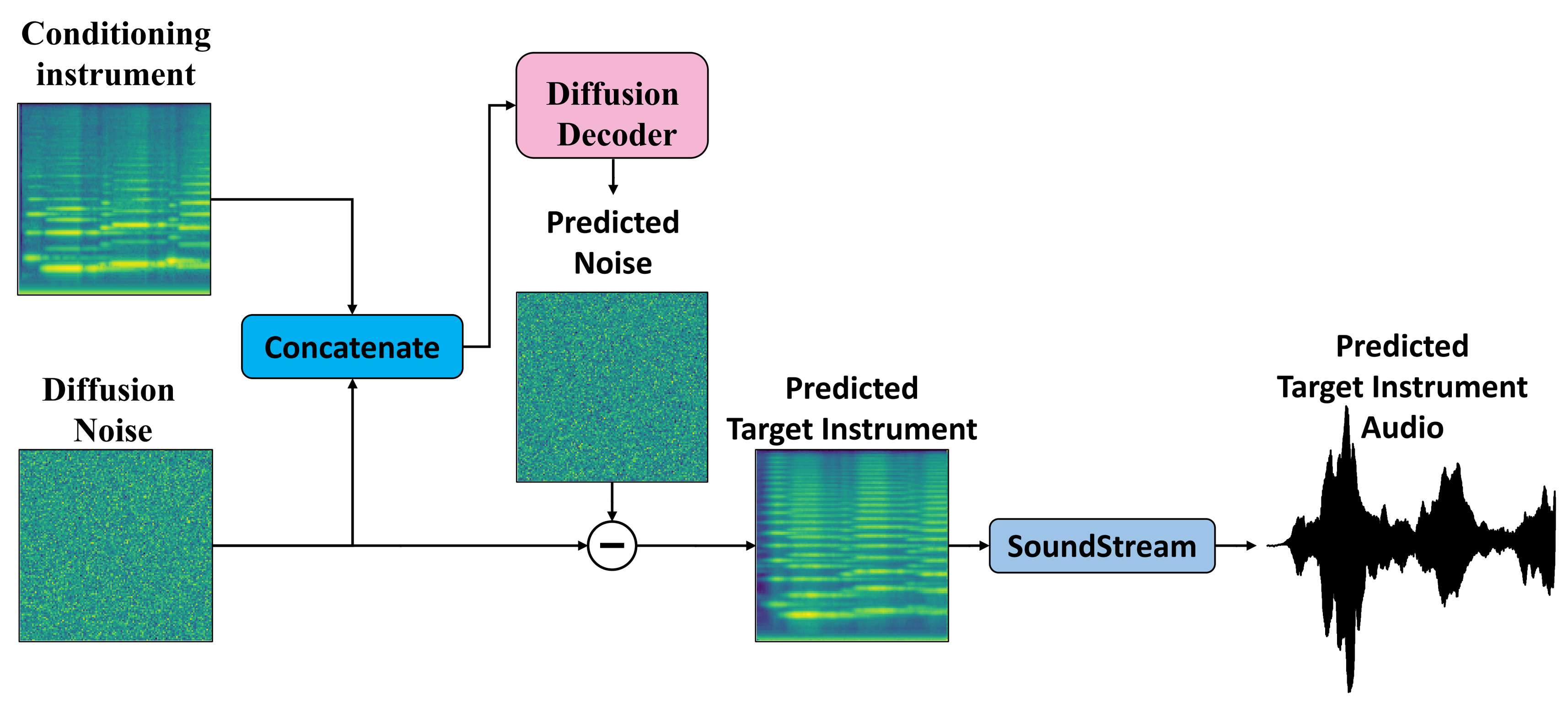
DiffTransfer

- Timbre transfer achieved through *conditional denoising diffusion implicit model*
- Log mel-scaled spectrograms converted from one timbre to another while keeping musical content
- Audio track reconstructed through pre-trained SoundStream Decoder[3]

Training



Inference



- Training procedure similar to image-to-image model *Palette*[4]: Conditioning instrument concatenated with noise
- At inference time only conditioning instrument is needed
- Model needs to be retrained if type of instruments are changed

Evaluation

- We use the StarNet dataset [5]
  - *Strings-Piano* and *Vibraphone-Clarinet* paired 16 kHz audio tracks
- We compare DiffTransfer with
  - Universal Network [6]: for single instrument timbre transfer
  - Music-STAR (*mixture-supervised*) model [7]: for multi-instrument timbre transfer
- We consider three timbre transfer tasks
  - ***Single***: only single instruments are converted
  - ***Single/mixed***: separate conversions of single instruments are mixed in order to create the desired mixture track
  - ***Mixture***: the mixture is directly converted
- **Training Procedure**
  - DiffTransfer trained for 5000 epochs using batch size 16 with AdamW optimizer
  - 6 models trained: *vibraphone to piano, piano to vibraphone, clarinet to strings, strings to clarinet, vibraphone/clarinet to piano/strings* and *piano/strings to vibraphone/clarinet*.

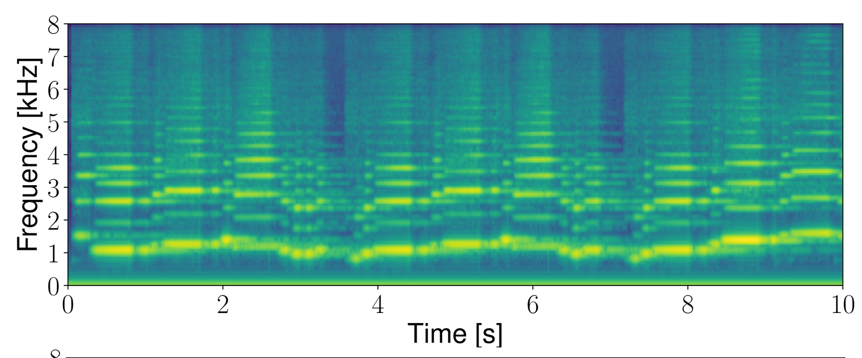
- **Objective Evaluation**
  - *Fréchet Audio Distance (FAD)*[8]: reference-free metric for music enhancement algorithms, measures perceptual similarity between the generated audios with respect to the ground truth one
  - *Jaccard Distance*: perceptual similarity between the generated audios with respect to the ground truth one

- **Subjective Evaluation**
  - Listening test, 18 human participants, split into two parts
    - Single Instrument timbre transfer
    - Multiple instrument timbre transfer
  - Conditions rated in terms of similarity with respect to reference track on a 1 (bad) - 5 (Excellent) Likert scale

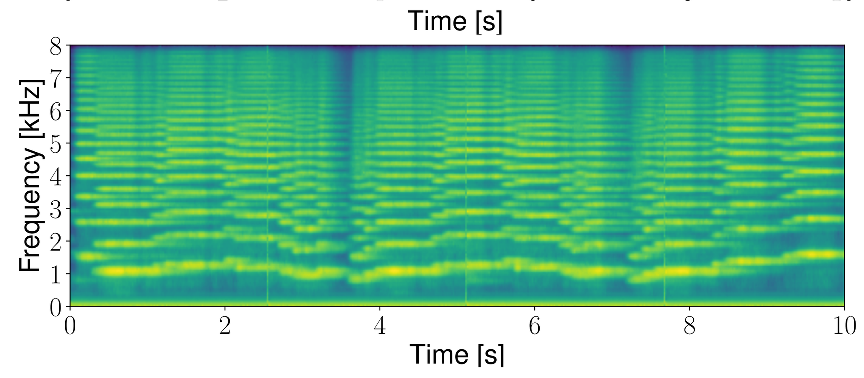
Objective Evaluation		
Method	**FAD ↓**	**JD ↓**
Universal Network (single)	7.09	0.53
DiffTransfer (single)	2.58	0.28
Universal Network (single/mixed)	10.47	0.64
DiffTransfer (single/mixed)	4.73	0.46
Music-STAR (mixture)	8.93	0.57
DiffTransfer (mixture)	4.37	0.38

Subjective Evaluation	
Method	**Similarity**
Universal Network (single)	1.82
DiffTransfer (single)	3.68
Universal Network (single/mixed)	1.69
DiffTransfer (single/mixed)	3.78
Music-STAR (mixture)	2.89
DiffTransfer (mixture)	3.80

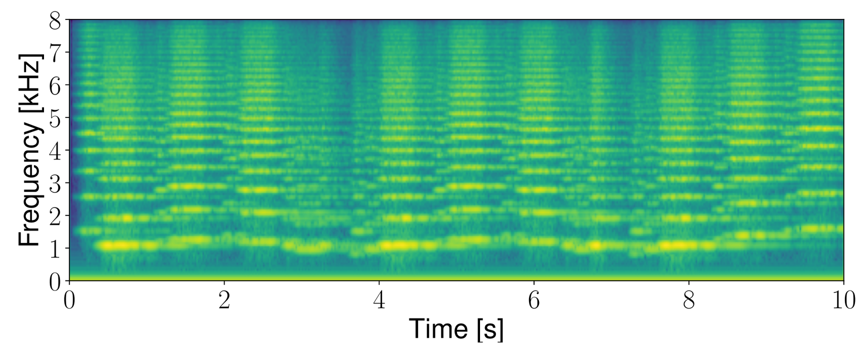
- **Input:** Clarinet



- **Output (DiffTransfer):** Strings



- **Ground Truth:** Clarinet



References

- [1] Colonel, Joseph T., and Sam Keene. "Conditioning autoencoder latent spaces for real-time timbre interpolation and synthesis." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.
- [2] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in International Conference on Learning Representations, 2021.
- [3] Zeghidour, Neil, et al. "Soundstream: An end-to-end neural audio codec." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021): 495-507.
- [4] Saharia, Chitwan, et al. "Palette: Image-to-image diffusion models." ACM SIGGRAPH 2022 Conference Proceedings. 2022.
- [5] M. Alinoori and V. Tzerpos, "Starnet," Aug. 2022. [Online]. Available: https://zenodo.org/record/6917099
- [6] A.P.Noam Mor, Lior Wold and Y.Taigman, "A universal music translation network," in International Conference on Learning Representations (ICLR), 2019.
- [7] M. Alinoori and V. Tzerpos, "Music-star: a style translation system for audio-based re-instrumentation," in 21st International Society for Music Information Retrieval (ISMIR2022), 2022.
- [8] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms." in INTER- SPEECH, 2019, pp. 2350–2354.

Scan for GitHub + Listening Examples!