ICML
International Conference
On Machine Learning

Unaligned Supervision for Automatic Music Transcription in The Wild

Ben Maman (Tel Aviv University)
Amit Bermano (Tel Aviv University)

Unaligned Supervision

Beethoven - Symphony No. 5 (Proms 2012)

11M views • 9 years ago

Mandetriens

Prom 12: **Beethoven** Cycle -- **Symphonies** Nos. **5** & 6 **Beethoven** - **Symphony No. 5** in C minor, Op. 67 1 - Allegro con brio 2 ...

https://www.youtube.com/watch?v=jv2WJMVPQi8&ab_channel=Mandetriens

Ludwig van Beethoven 5th Symphony By Herbert Von Karajan

304K views • 2 years ago

Champ

Quinta sinfonía de **Beethoven** dirigida por Herbert von Karajan.

https://www.youtube.com/watch?v=D-_wqx76mpc&ab_channel=Champ

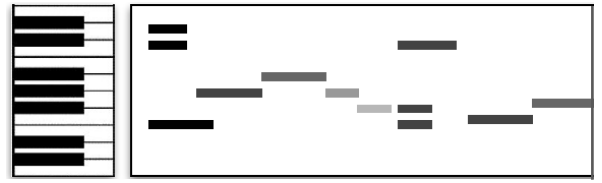
Beethoven - Symphony No. 5 - Iván Fischer | Concertgebouworkest

323K views • 2 years ago

Concertgebouworkest

Beethoven - **Symphony No. 5** - Iván Fischer | Concertgebouworkest Watch more videos free of charge on our...

https://www.youtube.com/results?search_query=ludwig+van+beethoven+5th+symphony+ivan+fischer



Beethoven Symphony 5.mid

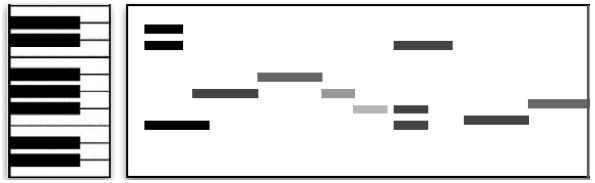
Image from [Hawthorne et al 2019]

Unaligned Supervision

Easily attainable data -> high quality supervision

https://www.youtube.com/watch?v=jv2WJMVPQi8&ab_channel=Mandetriens

+

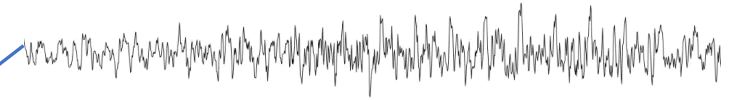


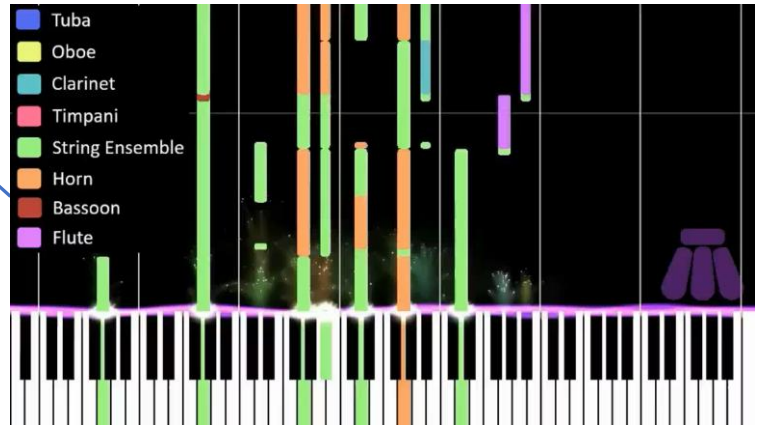
Beethoven Symphony 5.mid

Image from [Hawthorne et al 2019]

High-Quality Transcription

Brahms - Hungarian Dance No.5 - Hungarian Symphony Orchestra Budapest
https://www.youtube.com/watch?v=Nzo3atXtm54&ab_channel=MelosKonzerte

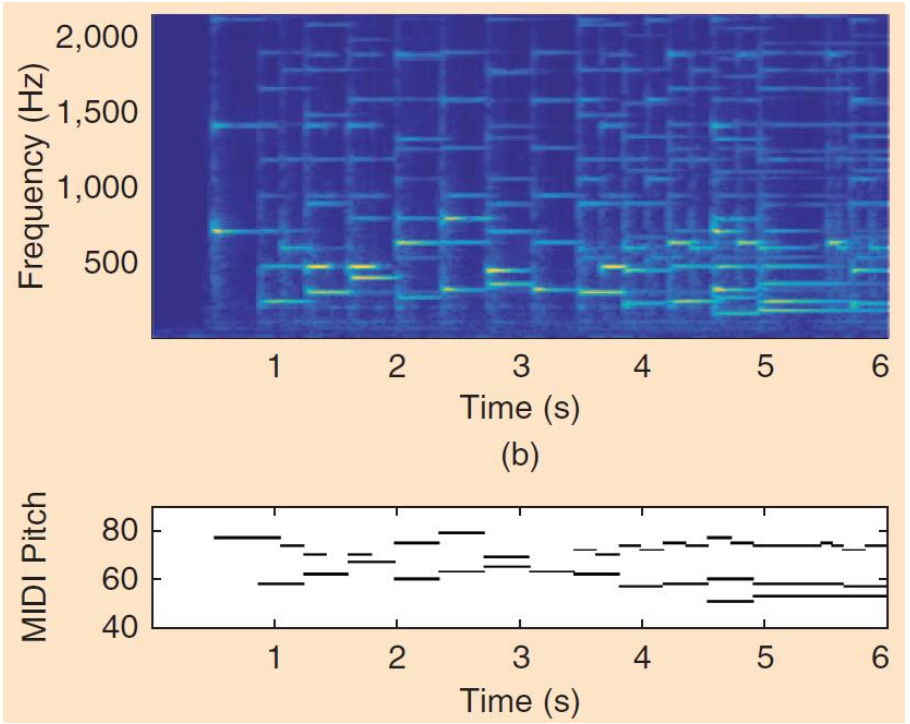




Automatic Music Transcription

- Input: Audio (spectrogram)
- Output: Note events (MIDI / Piano Roll)

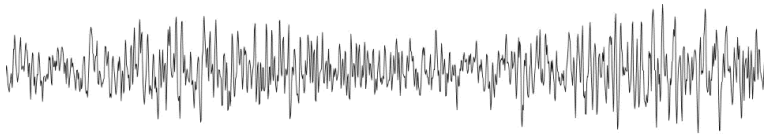
- Challenges –
  - Timing – note begin / end
  - Polyphony
  - Multi-Instrument
  - Fundamental vs. Partial frequencies

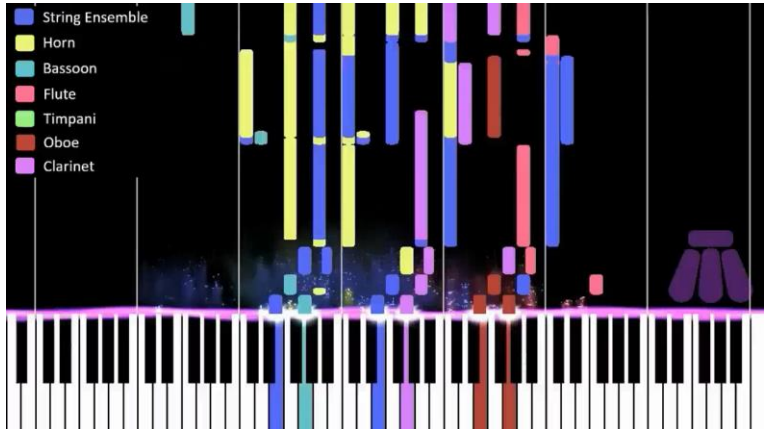


Benetos et al.,
Automatic Music Transcription: An Overview. IEEE Signal Process. Mag. (2019)

Bizet : "Carmen" Overture conducted by Myung-Whun Chung

https://www.youtube.com/watch?v=jL-Csf1pNCI&ab_channel=FranceMusique

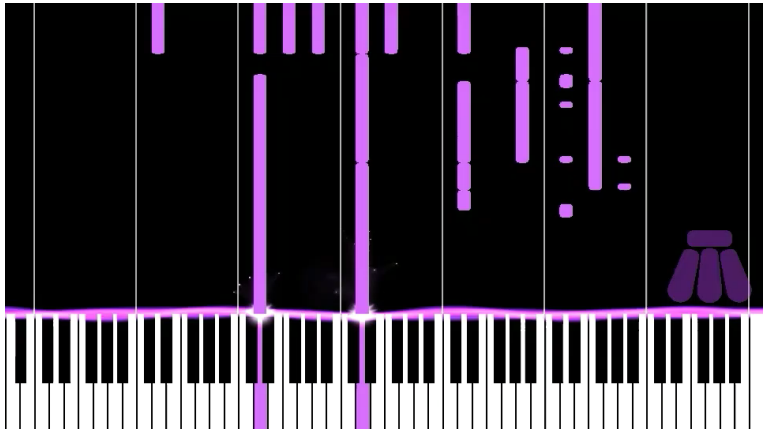




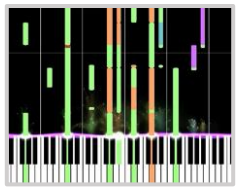
Data?

ABBA Gimme! Gimme! Gimme!

https://www.youtube.com/watch?v=JWay7CDEyAI&ab_channel=CraigGagn%C3%A9



Unaligned Supervision



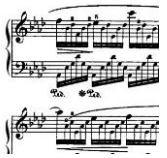
Annotated Data

Weak Detector

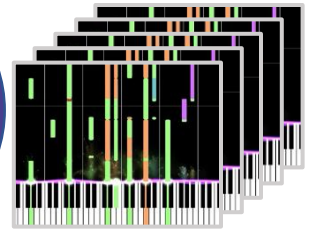
Alignment

Unaligned Data

+



Labeled Data



Piano - MAESTRO, MAPS

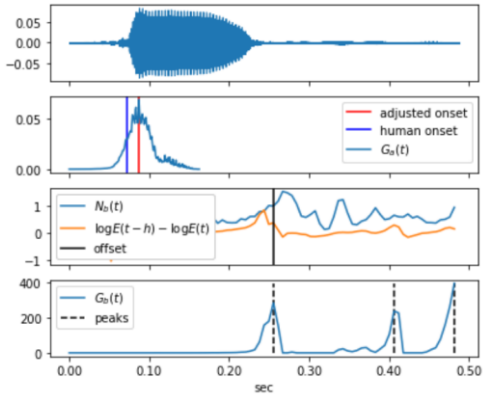
Hawthorne et al., **"Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset"**, ICLR 2019

- Fully supervised datasets (Yamaha Disklavier)
- MAPS – train synth, test real, MAESTRO – 130 hours real

https://www.youtube.com/watch?v=hBcdyFseDpM&ab_channel=AlamoMusicCenter

Guitar-Set

Xi et al., **"GUITARSET: A Dataset for Guitar Transcription"**, ISMIR 2018
3 hours acoustic guitar, semi-automatic annotation



Other/Multiple Instruments / Genres?

Only Poorly supervised data - MusicNet
Thickstun et al., **"Learning Features of Music From Scratch"**, ICLR 2017

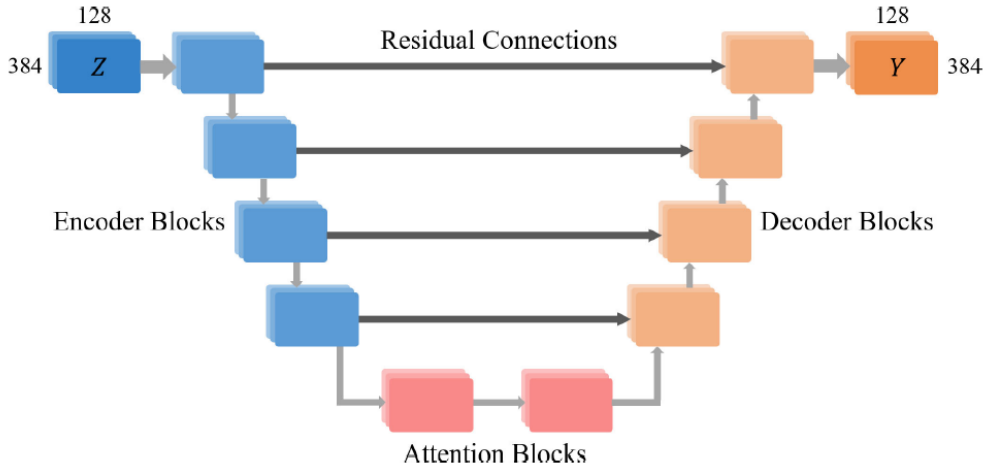
https://en.wikipedia.org/wiki/Classical_music#/media/File:Baschenis_-_Musical_Instruments.jpg

https://en.wikipedia.org/wiki/Jazz#/media/File:Charlie_Parker,_Tommy_Potter,_Miles_Davis,_Max_Roach_(Gottlieb_06941).jpg

https://en.wikipedia.org/wiki/Rock_music#/media/File:Elvis_Presley_promoting_Jailhouse_Rock.jpg

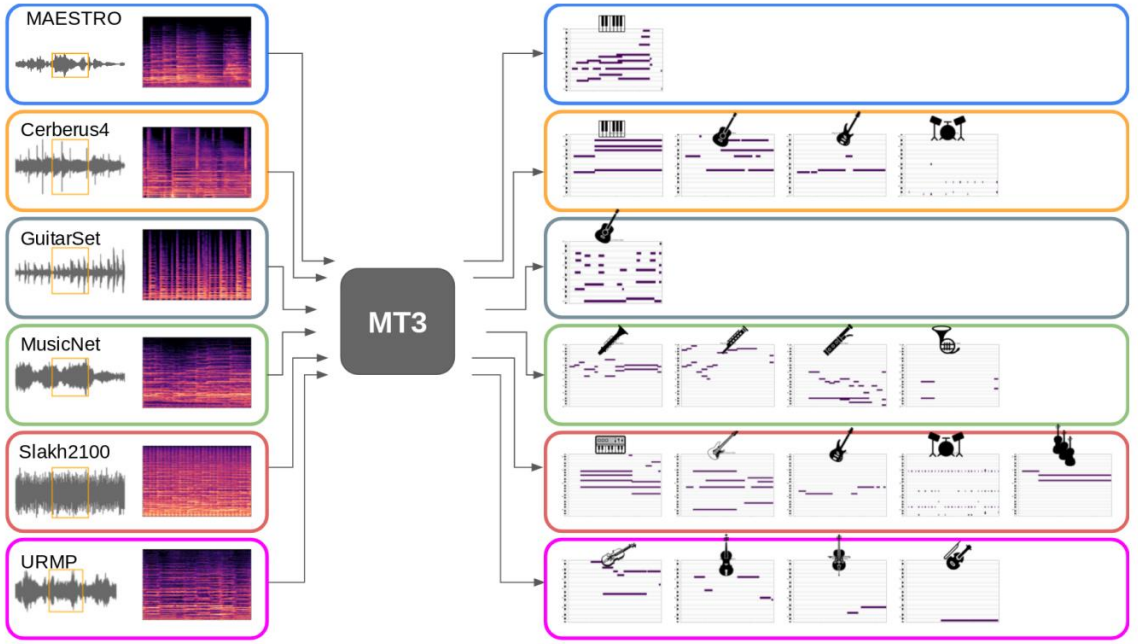
Omnizart

Wu et al., **Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation**
IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020



MT3

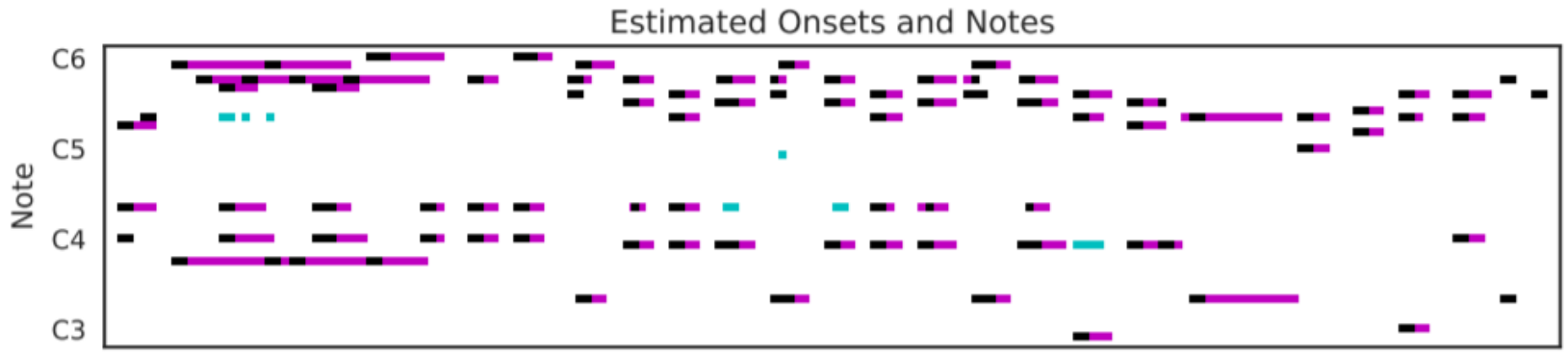
Gardner et al., **"MT3: Multi-Task Multitrack Music Transcription"**
ICLR 2022

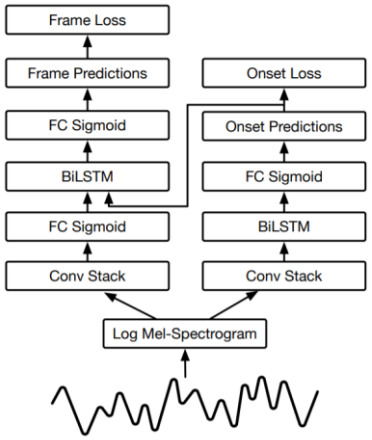


Onsets and Frames / Maestro

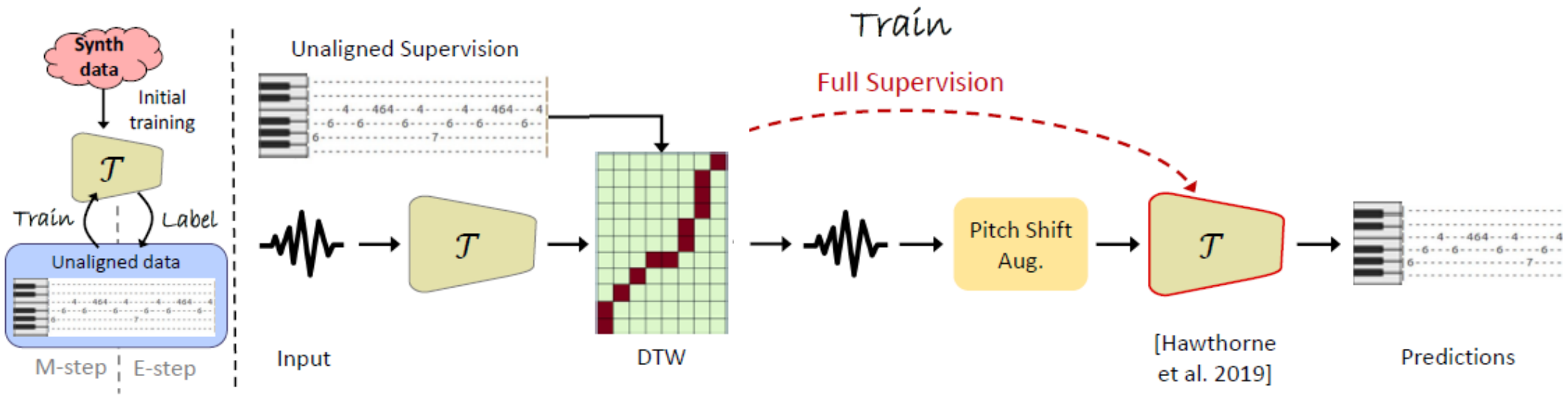
- Hawthorne et al., **“Onsets and Frames: Dual-Objective Piano Transcription”**, ISMIR 2017
- Hawthorne et al., **“Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset”**, ICLR 2019

- **Onsets =** note beginnings, **Frames=** note presence / duration





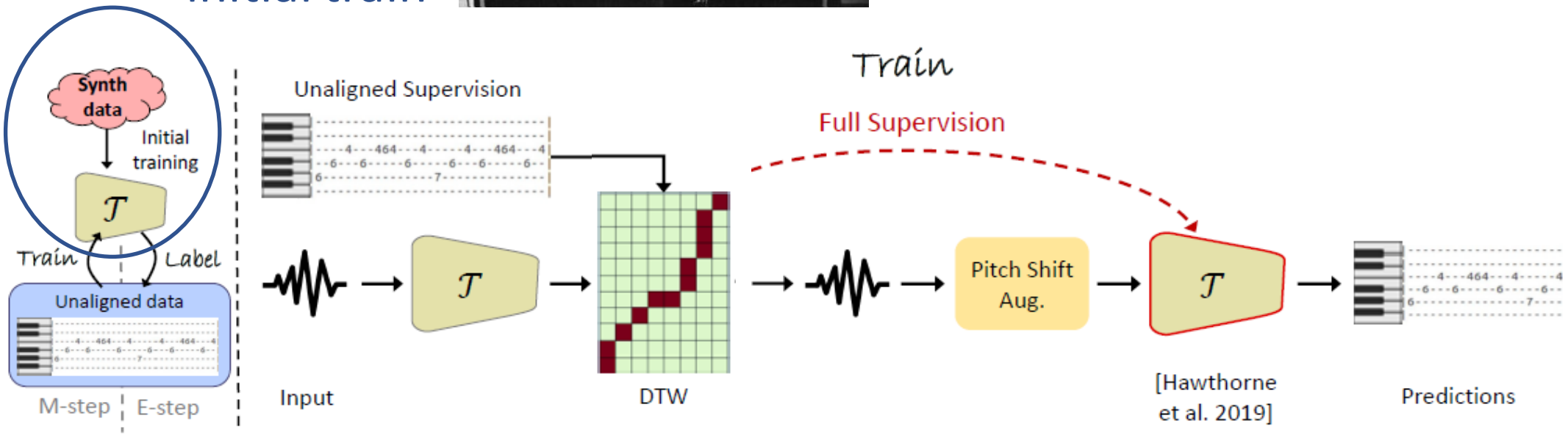
Pipeline


Synth data
Initial training
Train
Label
Unaligned data
M-step
E-step
Unaligned Supervision
Input
DTW
Train
Full Supervision
Pitch Shift Aug.
[Hawthorne et al. 2019]
Predictions

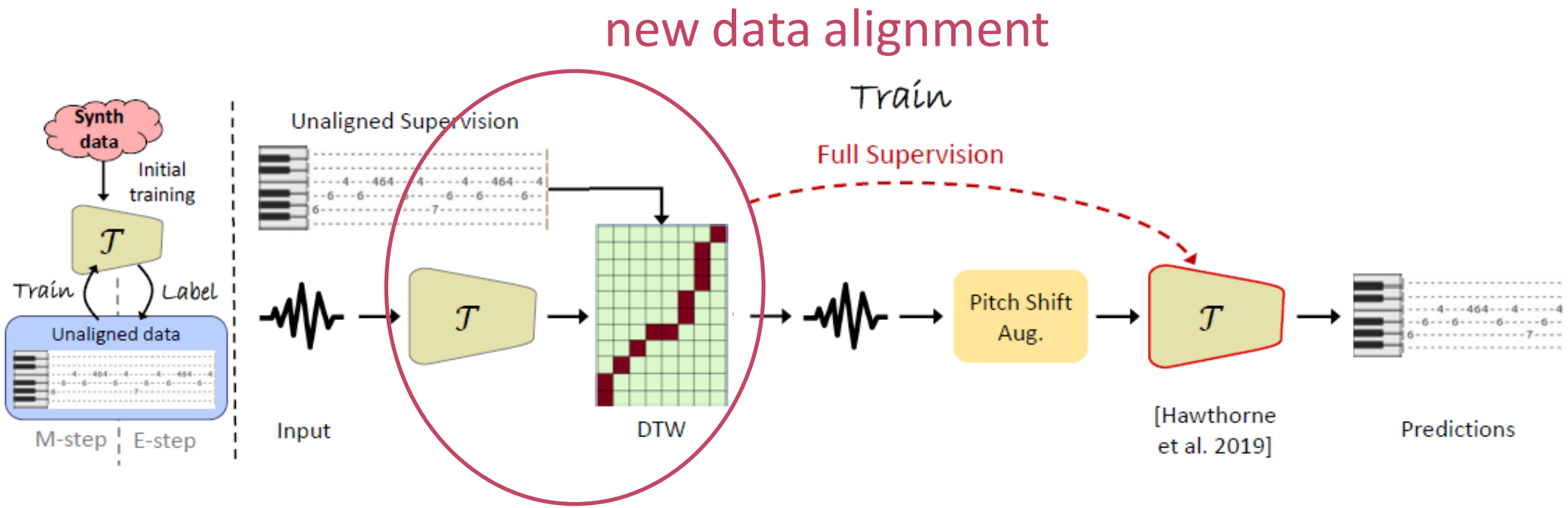
Pipeline

https://en.wikipedia.org/wiki/Synthesizer#/media/File:Bob_Moog3.jpg

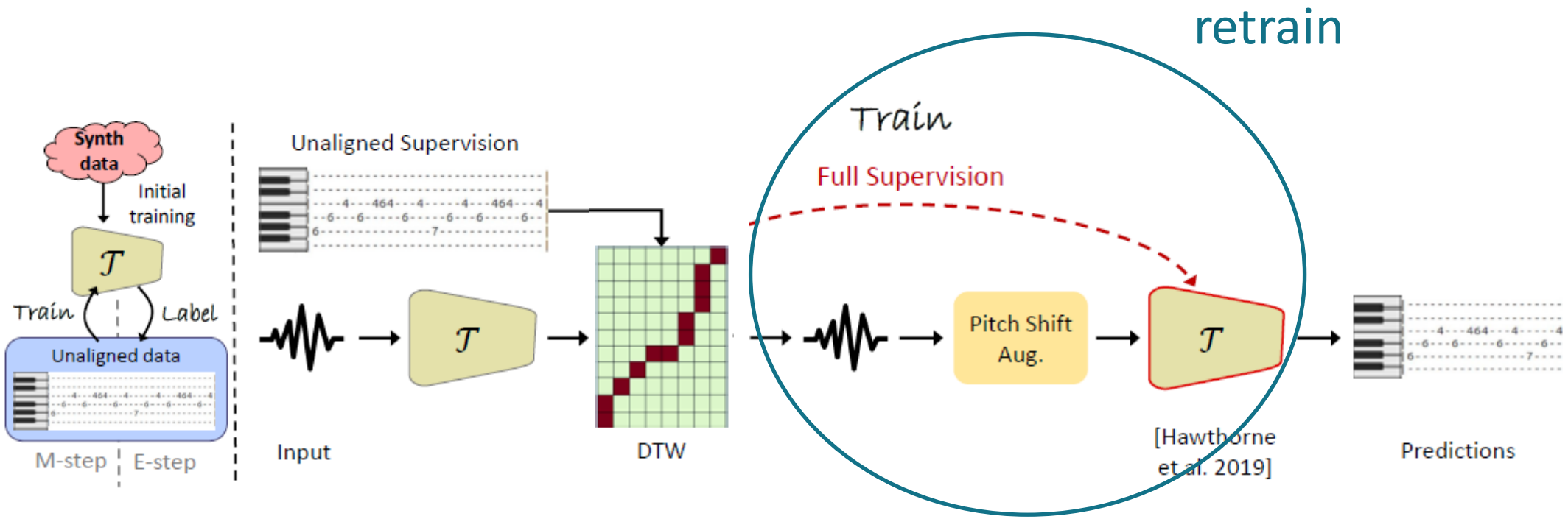
initial train



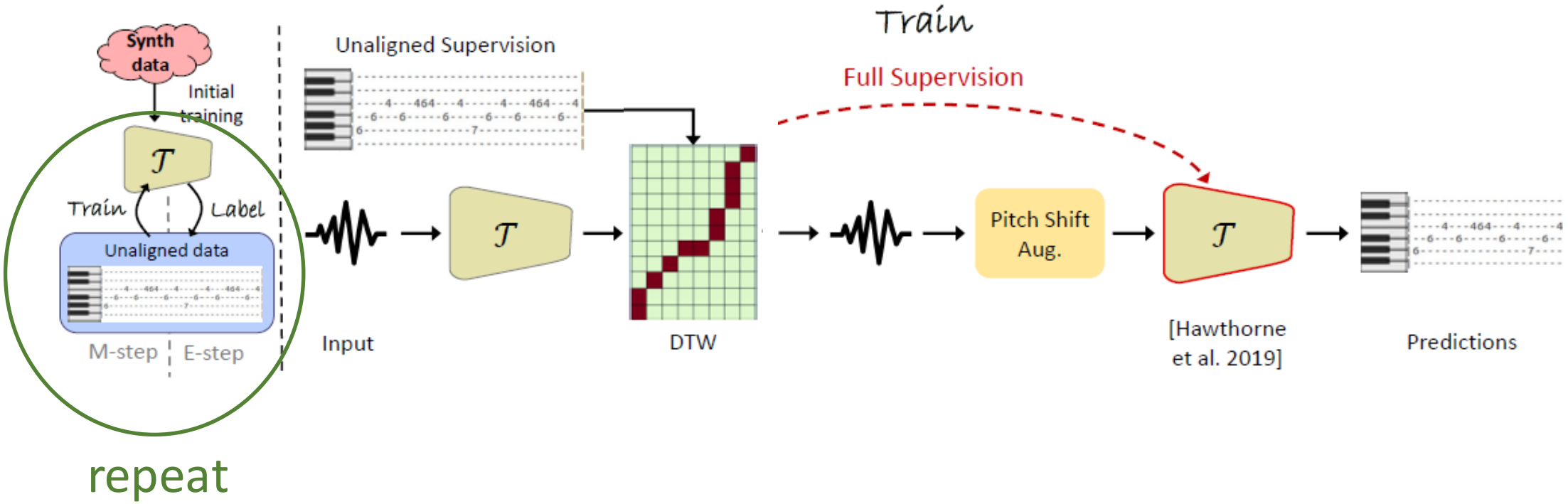
Pipeline


new data alignment
Synth data
Initial training
Train
Label
Unaligned data
M-step
E-step
Unaligned Supervision
Input
DTW
Train
Full Supervision
Pitch Shift Aug.
[Hawthorne et al. 2019]
Predictions

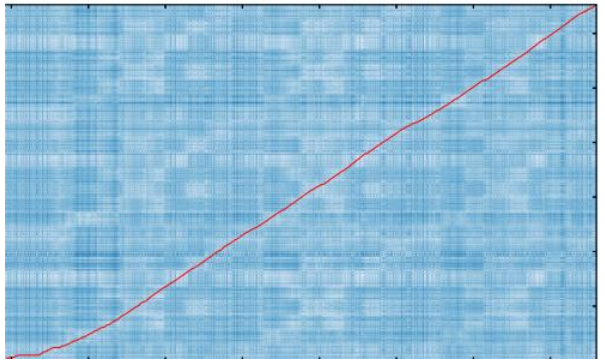
Pipeline


retrain
Synth data
Initial training
Train
Label
Unaligned data
M-step
E-step
Unaligned Supervision
Input
DTW
Train
Full Supervision
Pitch Shift Aug.
[Hawthorne et al. 2019]
Predictions

Pipeline


Synth data
Initial training
Train
Label
Unaligned data
M-step
E-step
repeat
Unaligned Supervision
Input
DTW
Train
Full Supervision
Pitch Shift Aug.
[Hawthorne et al. 2019]
Predictions

Alignment – Dynamic Time Warping



- MusicNet (Thickstun et al. 2016) – **spectral features DTW**

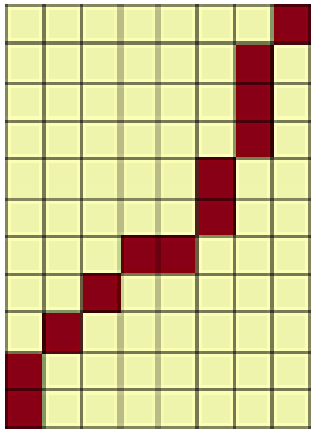
- Ours - DTW with initial network predicted **onset heatmap**

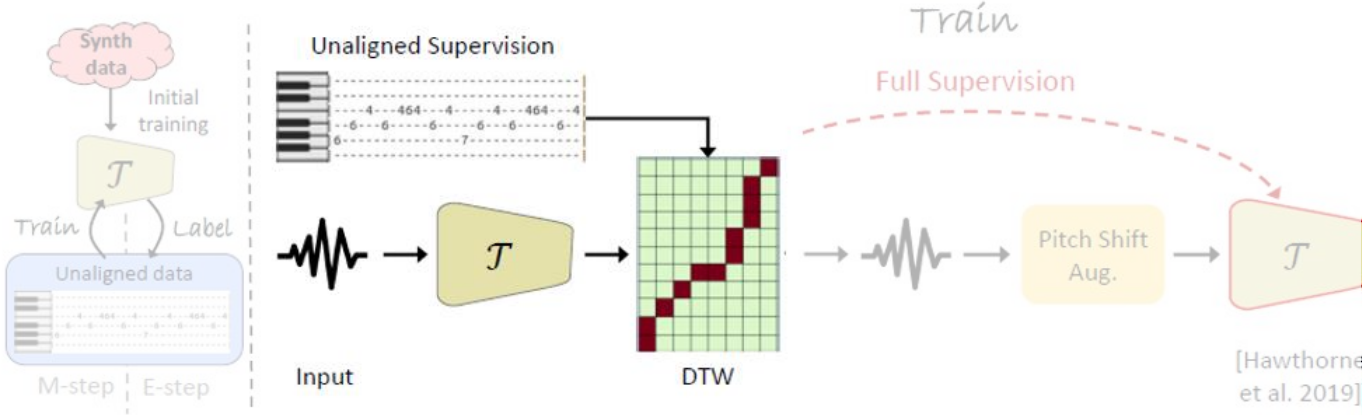
  onset=note beginning

$$\begin{aligned} \underset{t \in \mathbb{Z}^n}{\text{minimize}} \quad & \sum_{i=1}^{n} C(\mathbf{X}_{t_i}, \mathbf{Y}_i) \\ \text{subject to} \quad & t_0 = 0, \\ & t_n = m, \\ & t_i \leq t_j \qquad \text{if } i < j. \end{aligned}$$

- Local max => *precise* onset timing

✓ Support mismatches





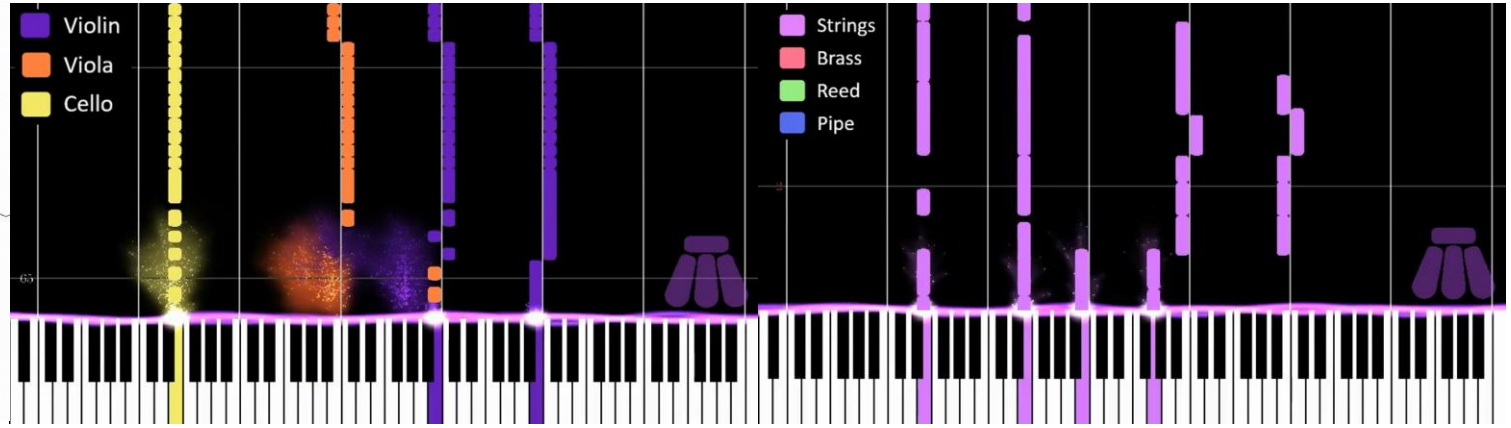
Data

Self-collected

Instrument	Length (Hours)
Piano	13:27:20
Harpsichord	6:20:37
Harpsichord & Strings	3:53:21
Harpsichord & Flute	1:02:18
Guitar	4:46:21
Lute	0:19:21
Violin	2:11:49
Cello	3:24:43
Flute	0:09:15
Organ	2:37:10
Orchestra	25:56:52
Orchestra & Piano	7:54:05
Orchestra & Choir	1:49:47
All	73:52:59
MusicNet	33:43:07
All, with MusicNet	107:36:06

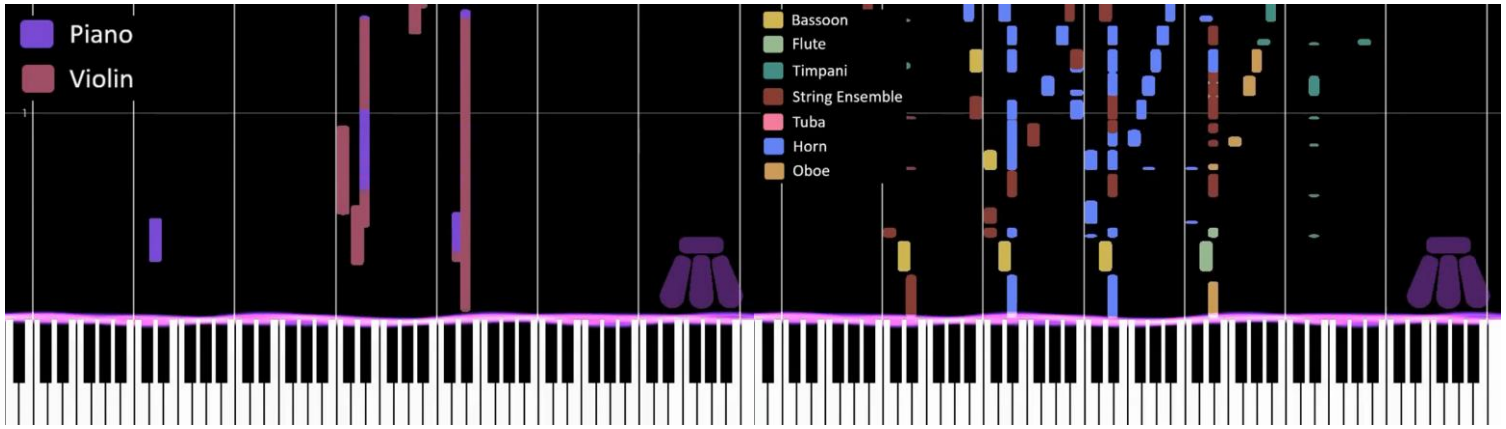
Qualitative Comparisons

Orquesta Alhambra - Rossini The Barber Of Seville Overture
https://www.youtube.com/watch?v=OloXRhesab0&t=2s&ab_channel=ClassicalMusicOnly



MT3 Ours

US Marine Band - Sousa's "The Stars and Stripes Forever“
https://www.youtube.com/watch?v=a-7XWhyvIpE



Omnizart Ours

Quantitative Evaluation

Main Results

Train \ Test	MAESTRO		MAPS		GuitarSet	
Supervised / Same dataset	Note	Frame	Note	Frame	Note	Frame
Hawthorne et al. (2019) (MAESTRO)	95.3	**90.2**	**86.4**	**84.9**	-	-
Gardner et al. (2021) (Mixture)	**96.0**	88.0	-	-	**90.0**	**89.0**
Weakly-supervised / Cross-dataset						
Gardner et al. (2021) ZS	28.0	60.0	-	-	32.0	58.0
Synth	83.8	74.7	79.1	76.6	68.4	72.9
Fine-tuned from Synth:						
MusicNet (original labels)	57.5	57.9	53.4	74.3	10.0	57.2
MusicNet$_{EM}$ (ours)	**89.7**	**76.0**	**87.3**	**79.6**	**82.9**	**81.6**

- Note-level
  Onset (note beginning) within 50ms
- Frame-level
  Note presence / duration

Train \ Test	MusicNet **Strings**		MusicNet **Wind**	
	Note	Frame	Note	Frame
(Cheuk et al., 2021)	61.0	68.4	48.2	67.4
Synth	49.1	49.8	55.4	64.3
Fine-tuned from Synth:				
MusicNet	39.9	**69.4**	38.0	73.4
MusicNet$_{EM}$ (ours)	**63.9**	68.3	**60.9**	**74.2**
	(Gardner et al., 2021) split			
(Gardner et al., 2021)	50.0	68.0		

MusicNet original low quality test annotation

MusicNetEM Dataset

- Our new labels for MusicNet
- EM on small groups, split by ensemble
- Solo Piano, strings, wind
- Include test in EM

test instrument	note-with-inst. F1
Violin	87.3
Viola	61.1
Cello	79.9
Bassoon	78.0
Clarinet	86.8
Horn	75.0

- Train from scratch on all, excluding test

	note F1	note-with-inst. F1	frame F1	note-with-offset F1
MusicNetEM	91.4	88.1	82.5	71.4
MusicNetEM wind	88.5	79.9	83.1	65.0
MusicNetEM strings	89.1	85.5	82.6	77.7
MusicNetEM strings*	85.9	81.1	79.0	75.1

Future Work

- Human voice
- Single-piece training (train at inference)
- → Note-with-offset, supervised-level frame accuracy
- Adversarial training (MIDI manifold)
- Velocity (note intensity)
-  MIDI-conditioned synthesis
- Spectrogram reconstruction loss

Thank You!

More on our project page:

https://benadar293.github.io