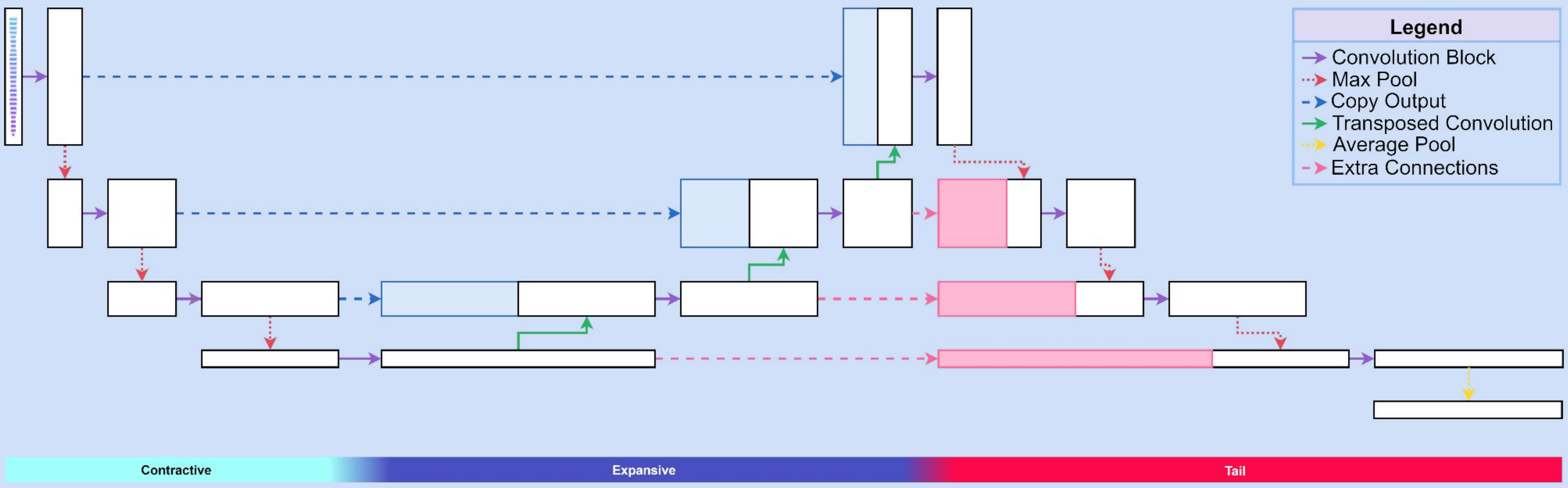
Hearing-Inspired AI Models Perform Better At Representing Music

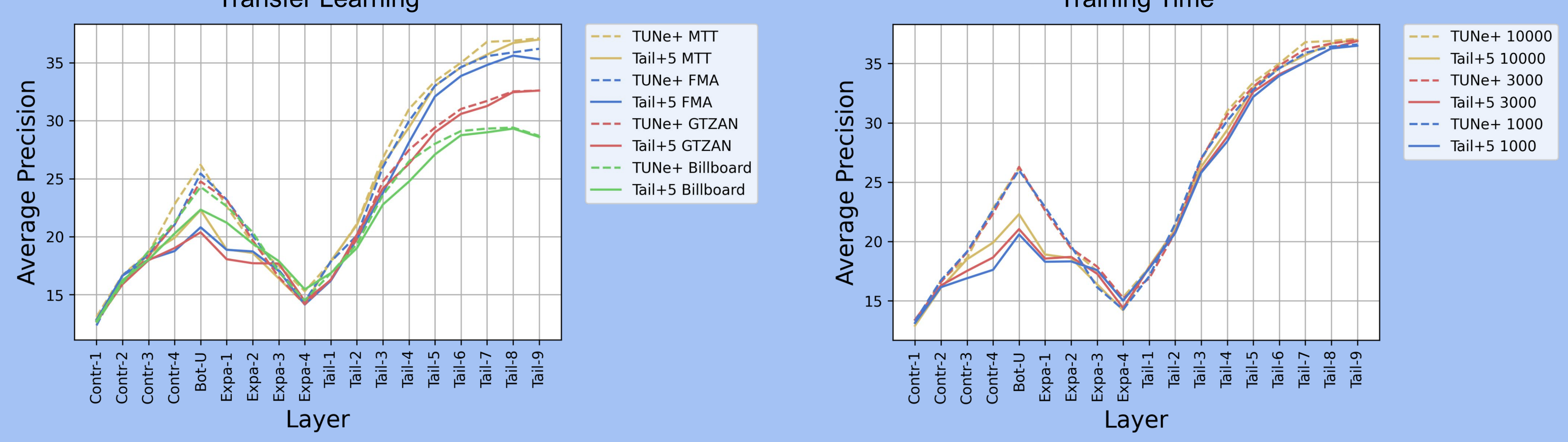
Marcel A. Vélez Vásquez, John Ashley Burgoyne
Music Cognition Group, Institute for Logic, Language and Computation, University of Amsterdam

Tailed U-Net architecture



We propose a novel architecture that outperforms self-supervised models and performs competitively compared to supervised models on downstream tasks.

Average precision across network layers



0. Problem statement

- Common representation learning architectures do not explicitly combine multi-scale features.
- U-Net architectures (Ronneberger et al., 2015) combine multi-scale features but the output is the size of the input ($\mathbb{R}^{l\times C}$) instead of representation size ($\mathbb{R}^{R}$).

1. The novel architecture intuition

Our architecture, which we call Tailed U-Net (TUNe), consists of three sections which can be easily shortened or lengthened:

- the **contractive path** extracts features at different scales;
- the **expansive path** combines features of different scales;
- the **tail path** maps the enriched signal to a latent space; and
- (for TUNe+) **extra connections** between the expansive and tail paths.

2. MTT training and probing

Variant	Supervised	Parameters	MTT_{AUC}	MTT_{AP}
TUNe Tail+5	**X**	**2.1 M**	**89.5**	**37.0**
TUNe+	**X**	**2.2 M**	**89.3**	**37.1**
CLMR	X	2.4 M	88.7	35.6
musicnn	✓	11.8 M	90.7	38.4

- At 10,000 epochs trained both TUNe variants outperform CLMR.

3. Out-of-domain training and MTT probing

Probing Variant	Training Data	MTT_{AUC}	MTT_{AP}
TUNe+	FMA	89.1	36.2
TUNe Tail +5	FMA	88.9	35.3
CLMR	FMA	86.2	30.6
TUNe+	GTZAN	87.2	32.6
TUNe Tail +5	GTZAN	86.9	32.6
CLMR	GTZAN	81.9	26.2
TUNe Tail +5	Billboard	84.7	28.6
TUNe+	Billboard	84.5	28.7
CLMR	Billboard	82.7	26.9

- For all three dataset both TUNe variants perform significantly better.

Additional Information

A. Architecture variant results for MTT training and probing

Variant	Filters	Parameters (M)	MTT_{AUC}	MTT_{AP}
Vanilla TUNe	34	2.4	87.7	33.0
TUNe Contractive+1	18	2.3	88.3	33.9
TUNe Contractive+2	9	2.1	88.6	34.6
TUNe Contractive+3	4	1.7	88.0	33.5
TUNe Expansive-1	34	2.3	87.6	33.0
TUNe Expansive-2	35	2.4	87.7	33.1
TUNe Expansive-3	38	2.3	87.6	33.0
TUNe Tail+1	28	2.3	88.2	34.4
TUNe Tail+2	19	2.3	88.7	35.2
TUNe Tail+3	15	2.3	89.1	36.5
TUNe Tail+4	13	2.3	89.2	36.6
TUNe Tail+5	11	2.1	89.2	36.5
TUNe CLMR-tail	10	2.5	89.4	36.7
TUNe+	11	2.2	89.2	36.6
Vanilla TUNe Small	11	0.4	86.8	31.9
TUNe+ Large	34	7.4	89.4	37.1
TUNe+ Smaller Rep	11	1.4	89.2	36.1

B. Training method and data

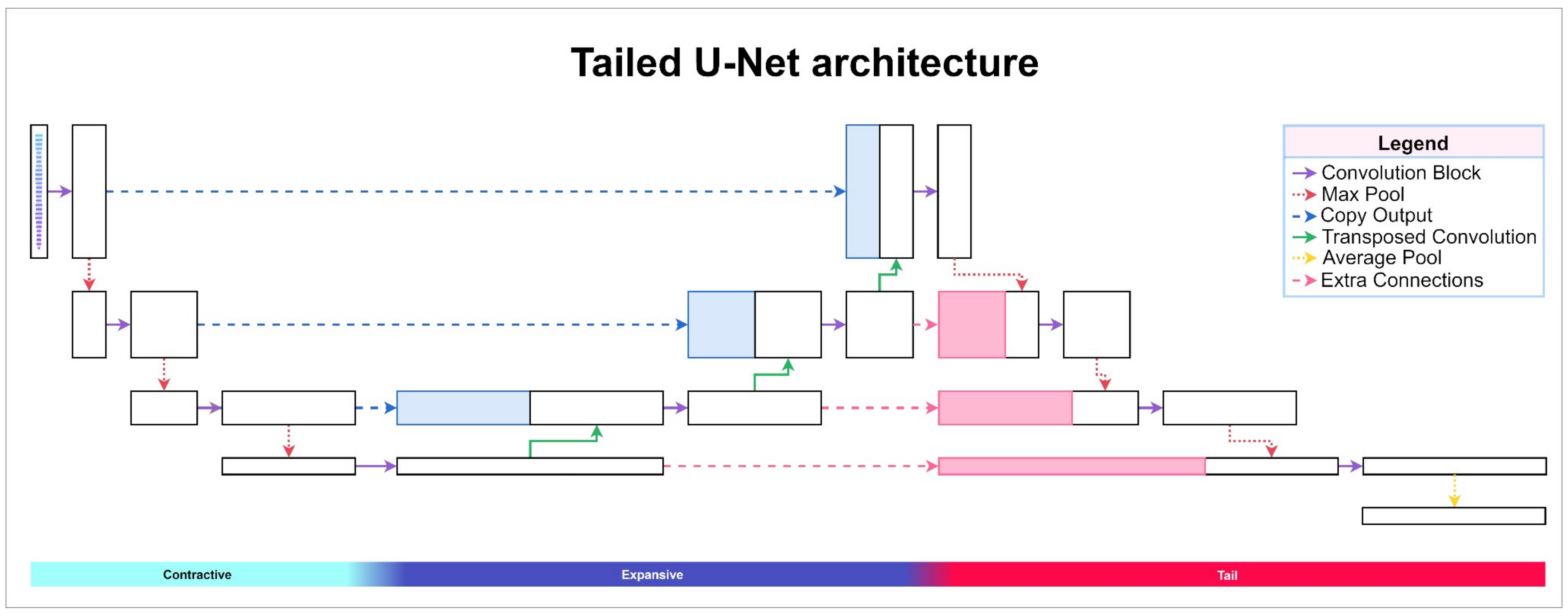
- The training method we used was CLMR (Spijkervet & Burgoyne, 2021).
- The datasets we used were:
  - MagnaTagATune Dataset (Law et al., 2009);
  - Free Music Archive (Defferrard, 2017);
  - GTZAN (Sturm, 2013); and
  - McGill Billboard (Burgoyne et al., 2011).

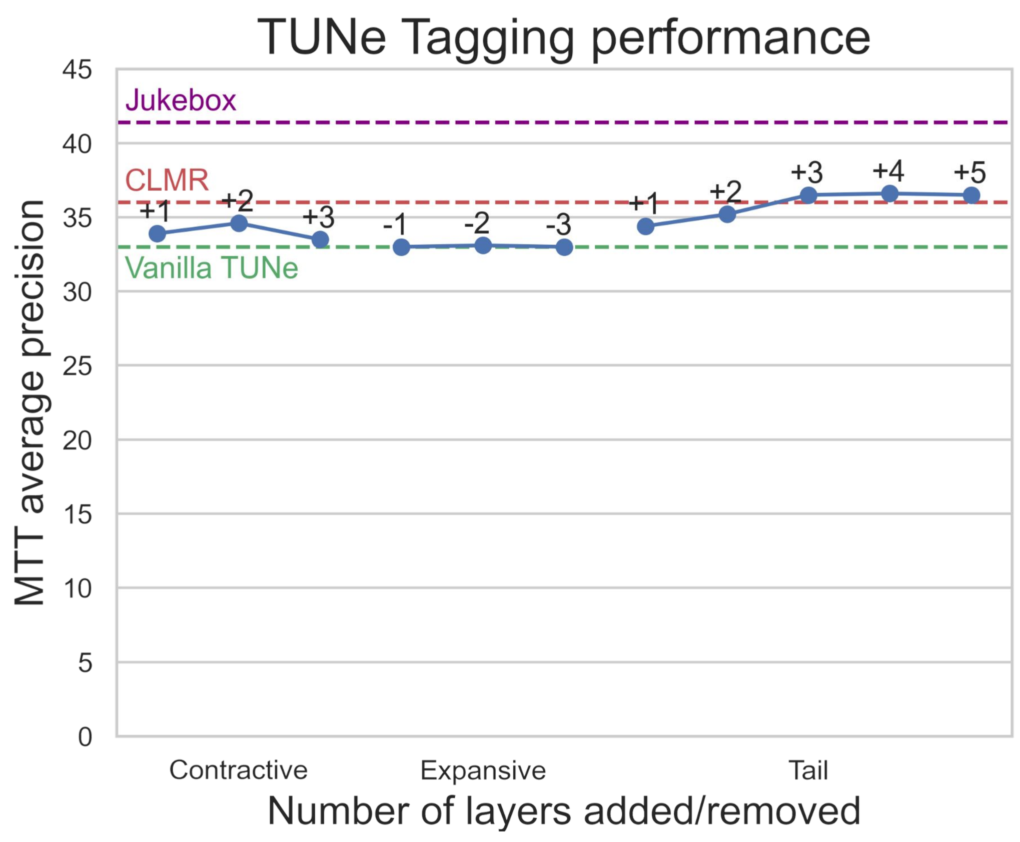
Github + Paper + Appendix

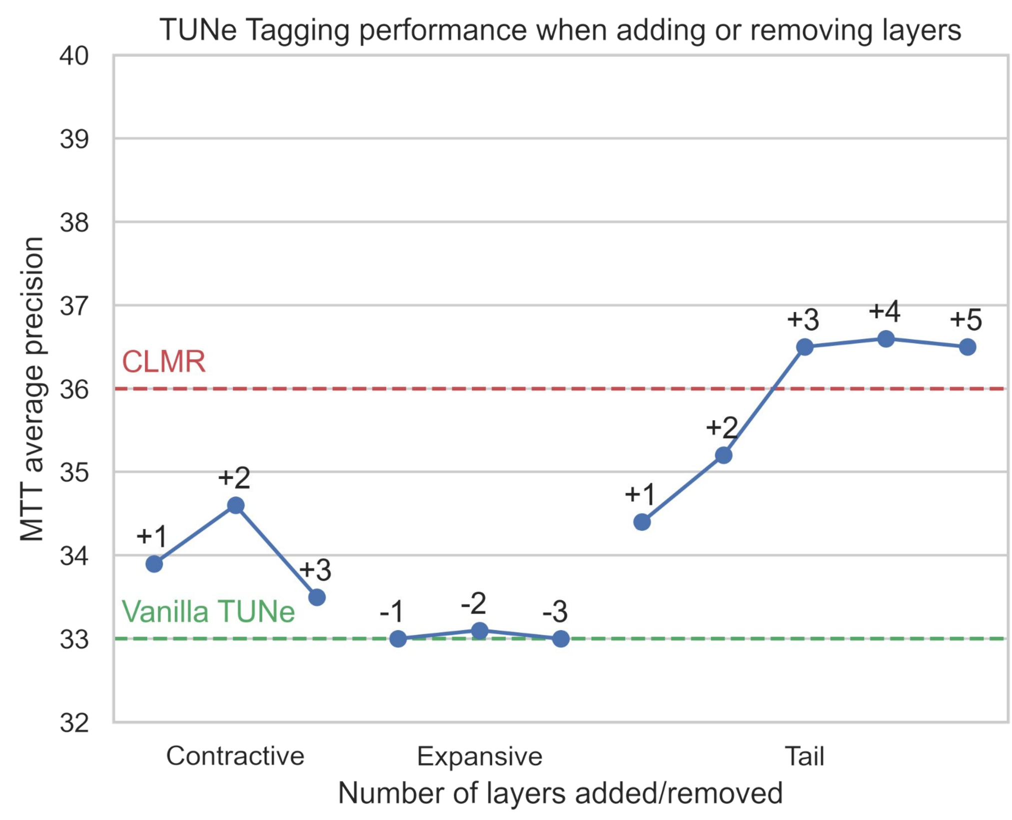
mcg.uva.nl

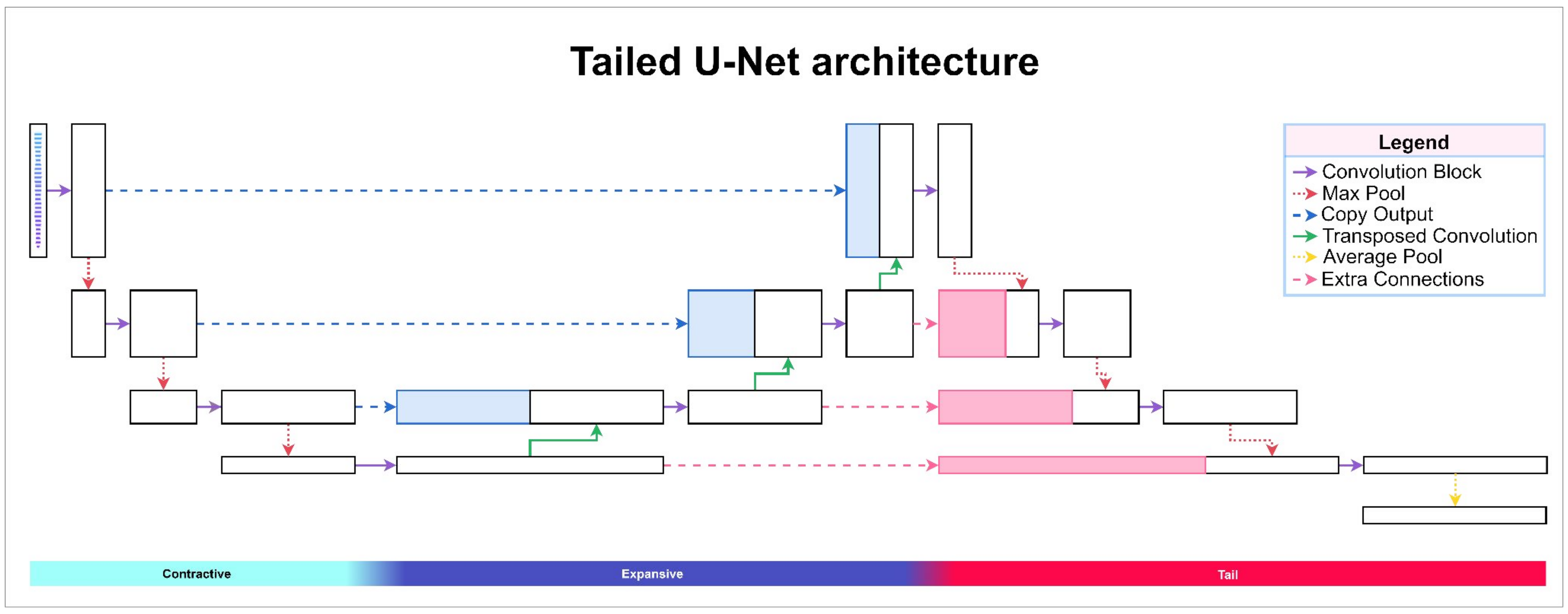
References

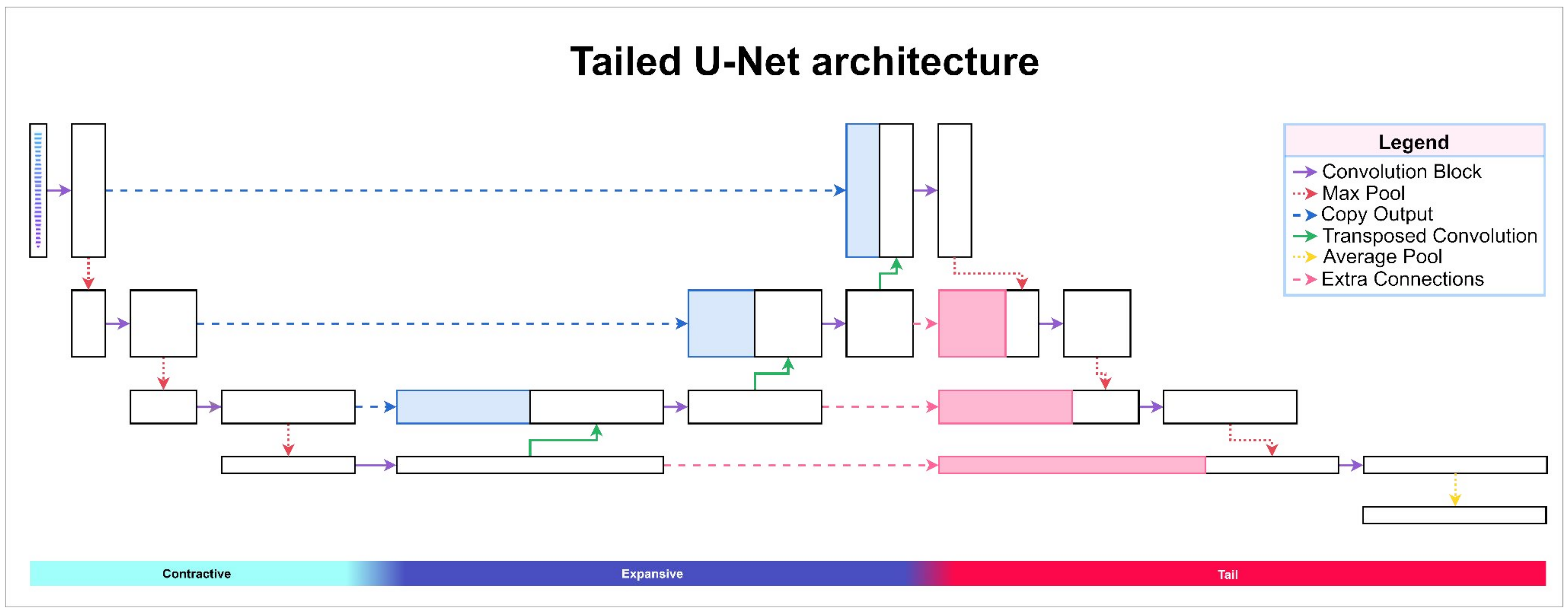
- O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 234–241.
- J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in Proceedings of the 22nd International Society for Music Information Retrieval Conference, 2021.
- J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," 2019.
- E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in Proceedings of the 10th International Society for Music Information Retrieval Conference, 2009, pp. 387–392.
- M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in Proceedings of the 18th International Society for Music Information Retrieval Conference, 2017.
- B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," arXiv:1306.1461, 2013.
- J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground truth set for audio chord recognition and music analysis," in Proceedings of the 12th International Society for Music Information Retrieval Conference, 2011, pp. 633–638.

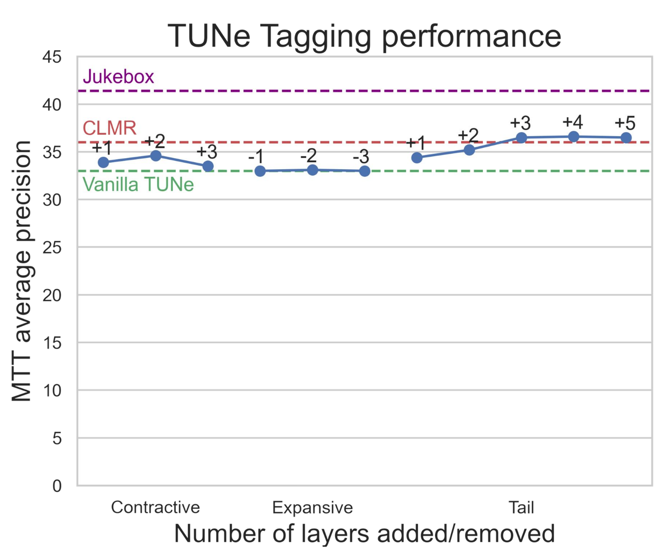

Tailed U-Net architecture
Legend
Convolution Block
Max Pool
Copy Output
Transposed Convolution
Average Pool
Extra Connections
Contractive
Expansive
Tail

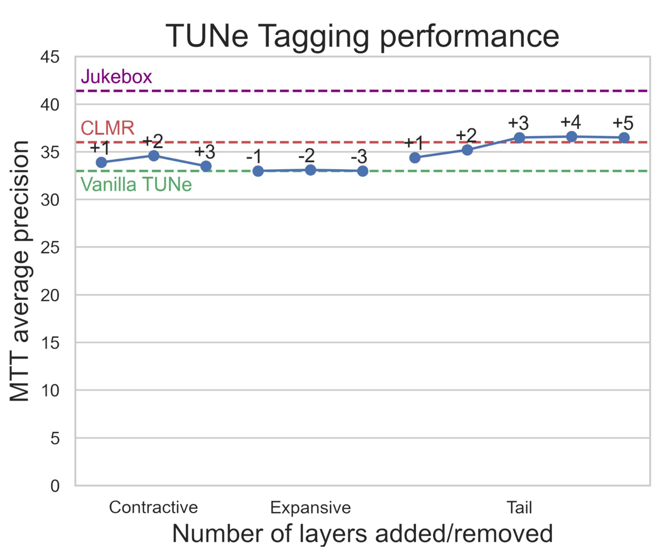

TUNe Tagging performance
MTT average precision
Jukebox
CLMR
Vanilla TUNe
+1
+2
+3
-1
-2
-3
+1
+2
+3
+4
+5
Contractive
Expansive
Tail
Number of layers added/removed

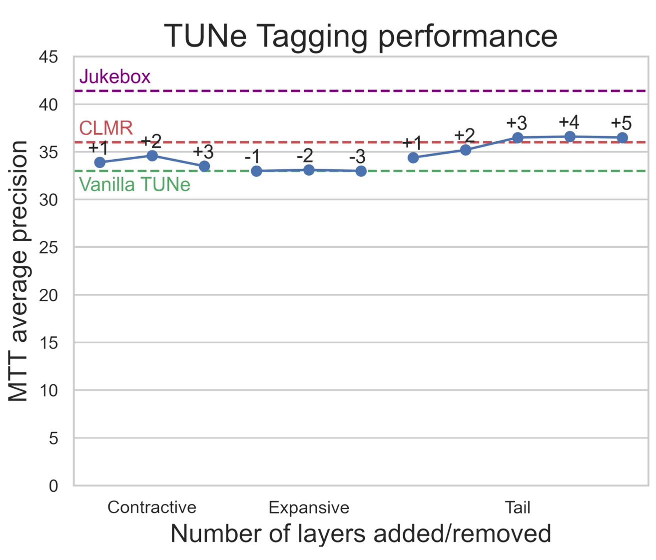

TUNe Tagging performance when adding or removing layers
MTT average precision
CLMR
Vanilla TUNe
+1
+2
+3
-1
-2
-3
+1
+2
+3
+4
+5
Contractive
Expansive
Tail
Number of layers added/removed

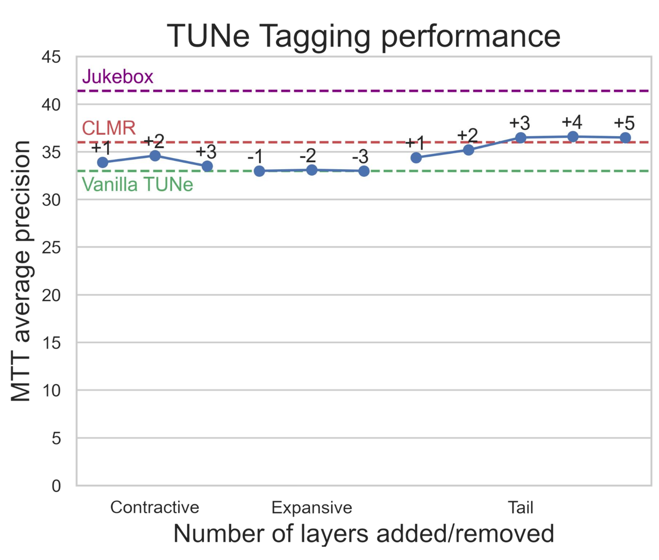

Tailed U-Net architecture
Legend
Convolution Block
Max Pool
Copy Output
Transposed Convolution
Average Pool
Extra Connections
Contractive
Expansive
Tail


Tailed U-Net architecture
Legend
Convolution Block
Max Pool
Copy Output
Transposed Convolution
Average Pool
Extra Connections
Contractive
Expansive
Tail


TUNe Tagging performance
MTT average precision
Jukebox
CLMR
Vanilla TUNe
Contractive
Expansive
Tail
Number of layers added/removed


TUNe Tagging performance
MTT average precision
Jukebox
CLMR
Vanilla TUNe
Contractive
Expansive
Tail
Number of layers added/removed


TUNe Tagging performance
MTT average precision
Jukebox
CLMR
Vanilla TUNe
Contractive
Expansive
Tail
Number of layers added/removed


TUNe Tagging performance
MTT average precision
Jukebox
CLMR
Vanilla TUNe
Contractive
Expansive
Tail
Number of layers added/removed