Reconstruction Meets Prediction: Complementary SSL Objectives for Music Transcription

Mary Pilataki, Matthias Mauch, and Simon Dixon

***Abstract*—Multi-instrument automatic music transcription (MIAMT) requires learning representations from limited labelled data. We present a controlled comparison of two self-supervised paradigms: masked spectrogram reconstruction with a masked autoencoder (MAE) and predictive learning with a joint embedding predictive architecture (JEPA). We use identical Transformer encoders, evaluating both transcription performance and latent space properties: geometric organisation, retrieval quality and representational capacity. On transcription, the two paradigms are broadly comparable, with task-specific advantages: JEPA performs better for monophonic wind instruments, MAE for dense polyphonic ensembles. Both substantially exceed the prior SSL MIAMT baseline, and on URMP both also exceed a recent state-of-the-art supervised system. In latent space, JEPA achieves better geometric organisation while MAE preserves higher representational capacity. Furthermore, MAE converges substantially faster under supervision. Initialising JEPA from MAE-pretrained weights consistently improves transcription, demonstrating that the two objectives are complementary rather than competing. Three key findings emerge. First, no single latent metric reliably predicts transcription performance. Second, MAE provides a richer initialisation for predictive pretraining than random. Third, reconstruction and prediction objectives reveal opposite capacity trajectories under supervision, a behaviour that persists regardless of initialisation.**

***Index Terms*—Music transcription, self-supervised learning, masked autoencoders, joint embedding predictive architectures.**

I. Introduction

MULTI-INSTRUMENT automatic music transcription (MIAMT) extracts note-level symbolic representations from polyphonic audio containing multiple instruments. It is fundamental to Music Information Retrieval (MIR), with applications in music education, interactive music systems and musicology. Despite deep learning advances, accurate transcription remains challenging due to overlapping harmonics, timbral diversity of source instruments and interrelated subtasks, including multi-pitch estimation (MPE), onset/offset detection and instrument recognition (IR) [1].

Received 19 March 2026; revised 1 June 2026; accepted 2 June 2026. Date of publication 16 June 2026; date of current version 8 July 2026. This work was supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music under Grant EP/S022694/1. The associate editor coordinating the review of this article and approving it for publication was Dr. Amin Zollanvari. *(Corresponding author: Mary Pilataki.)*

Mary Pilataki and Simon Dixon are with the Centre for Digital Music, Queen Mary University of London, E1 4NS London, U.K. (e-mail: m.pilataki-manika@qmul.ac.uk; s.e.dixon@qmul.ac.uk).

Matthias Mauch is with Apple, EC2M 4AG London, U.K. (e-mail: mmauch@apple.com).

Code and pretrained models will be released on GitHub https://github.com/marypilataki/amt-mae.

Digital Object Identifier 10.1109/LSP.2026.3704065

State-of-the-art systems rely on supervised learning with spectral features [2], [3], [4], [5], [6] but face a critical bottleneck: the scarcity of large-scale, high-quality annotated datasets. Manual note-level annotation requires expert musicians and is expensive, resulting in limited instrument and genre coverage. This constrains supervised approaches that depend on large labeled corpora.

Self-supervised learning (SSL) offers a promising alternative by learning representations from unlabeled data, with masked modeling being a dominant paradigm in the audio domain [7], [8], [9], [10], [11], [12]. Generative approaches like the Audio Masked Autoencoder (AudioMAE) [10], [13] and hybrid strategies that combine complementary discriminative and multimodal objectives [12], [14], [15], have shown strong downstream performance. Predictive frameworks based on Joint Embedding Predictive Architectures (JEPA), including Audio-JEPA [9] and music-specific variants [11], [16] and structured methods like MATPAC [17], [18] have shifted the focus towards learning directly in the representation space, bypassing signal reconstruction. However, how these self-supervised objectives shape the learned latent space and whether the resulting geometric and capacity differences translate into measurable performance differences remains unclear.

We hypothesise that the SSL objective fundamentally shapes both representation geometry and representational capacity, and that analysing these properties can reveal complementary strengths that transcription scores alone cannot expose. To test this, we present a controlled comparison of two SSL frameworks for MIAMT based on JEPA [8], [9], [11] and AudioMAE [10]. Both use identical Transformer-based spectrogram encoders and published hyperparameters, differing only in masking strategy and training objective. JEPA models high-level semantic relationships between contextual and target blocks directly in latent space, without signal reconstruction [8], while AudioMAE reconstructs missing spectrogram patches from the surrounding context, encouraging detailed spectral modelling [10], [19]. This controlled setup allows us to attribute any differences in latent structure and downstream performance to the SSL objective itself.

While MIR studies typically evaluate SSL models mainly through downstream performance alone, such *closed box* evaluation provides limited insight into representation structure. Inspired by recent encoder assessment frameworks [20], [21] and representation analyses used in vision [22] and audio [23] SSL, we extend evaluation beyond transcription metrics. Specifically, we measure geometric organisation via alignment and

1070-9908 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on August 25,2026 at 13:55:16 UTC from IEEE Xplore. Restrictions apply.

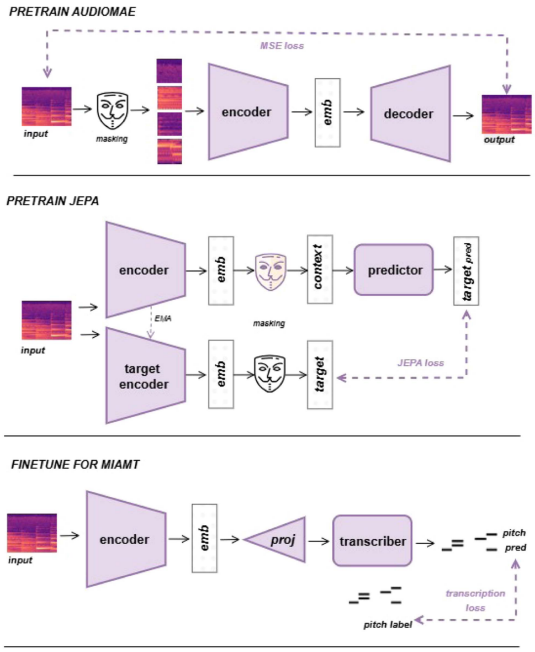


Fig. 1. SSL MIAMT framework: MIAMT-MAE pretraining (top), MIAMT-JEPA pretraining (middle) and the shared finetuning setup (bottom).

uniformity of the hypersphere [24], pitch-space neighbourhood structure via pitch locality, retrieval quality via kNN accuracy [22] and representational capacity by the RankMe score [21]. This analysis reveals how musical information is structured and utilised in the latent space, rather than relying on downstream scores alone.

Our main findings are as follows. Transcription performance is broadly comparable between the two models, with task-specific advantages. MIAMT-JEPA outperforms on monophonic wind instruments while MIAMT-MAE is stronger on dense polyphonic content (piano and ensemble). Both substantially exceed the prior SSL transcription baseline [11], and on URMP both also exceed a recent state-of-the-art supervised system. MAE-pretrained initialisation of JEPA (MIAMT-JEPA†) yields consistent gains across all tasks, demonstrating that the two objectives are complementary and that MAE provides a powerful initialisation for predictive pretraining. Additionally, MAE converges substantially faster than JEPA under supervised finetuning, an important practical consideration. Latent-space analysis reveals complementary trade-offs. MIAMT-JEPA achieves superior geometric organisation (uniformity, pitch locality) and retrieval quality, while MIAMT-MAE preserves higher representational capacity. Surprisingly, finetuning produces opposite capacity trajectories, with MAE expanding and JEPA reducing, a behaviour that persists regardless of initialisation. No single latent metric reliably predicts transcription performance, indicating that the two SSL objectives are best understood as complementary rather than competing.

II. Model Architectures

Both MIAMT-MAE and MIAMT-JEPA use identical Vision Transformer (ViT-B) encoders [25] adapted for spectrograms, differing only in pretraining modules (Fig. 1). Input log-mel spectrograms are divided into non-overlapping $[8 \times 8]$ patches and are linearly projected to a 768-dimensional embedding space with learnable positional encodings that retain temporal and spectral order. The sequence is processed by 12 Transformer blocks, each with 12 self-attention heads, a feed-forward layer with expansion ratio 4 and Layer Normalisation. The normalised output hidden states form the encoder representation, which is input to the transcription heads.

During pretraining, MIAMT-MAE uses an 8-block decoder (512-dim) to reconstruct masked patches [10], while MIAMT-JEPA uses a 6-block predictor (384-dim) to estimate target embeddings from an Exponential Moving Average (EMA) encoder in latent space [8]. Both pretraining modules are discarded at finetuning.

Following the MIAMT-JEPA baseline [11], both models share a shallow transcription head consisting of two hidden linear layers with sigmoid activations, producing pitch activity probabilities per instrument. The head accepts 128-dimensional inputs (one per mel bin) and produces an $[88 \times 11]$ matrix covering 88 pitches (A0-C8) and 11 instrument classes. Latent embeddings are projected into a spectrogram space via transposed convolution before being passed to the transcription head (Fig. 1, bottom).

III. Experimental Setup

A. Datasets and Input Features

Following [11], SSL pretraining uses GTZAN [26], the Violin Bach dataset [27], Mazurkas,[1] and the FrançoisLeduc guitar dataset [28], totalling approximately 159 hours of unlabeled audio. Supervised finetuning uses Slakh [29], MusicNet's EM version [30], [31], GuitarSet [32], URMP [33] and MAESTRO [34] with standard train-test splits from [2], [3].

Log-mel spectrograms are computed from 44.1 kHz mono audio using 128 mel bands, a 100 ms Hanning window, 10 ms hop size and a 20 Hz lower frequency bound, using torchaudio's *MelSpectrogram* transform [35]. This yields $[128 \times 128]$ representations corresponding to 1.28 s audio segments. Consecutive segments overlap by 25% (0.32 s stride). Spectrograms are normalised using dataset statistics [10].

B. Experiments

Both models are pretrained self-supervised for 300 epochs using their published configurations [8], [9], [10]. We then finetune each encoder jointly with the multi-instrument transcription head, discarding SSL-specific modules (decoder, predictor) and objectives and optimising only the transcription loss. No encoder layers are frozen during finetuning. Preliminary experiments showed that this transcription-loss-only finetuning setup outperforms frozen-encoder probing. In [11], three variants were compared, end-to-end training, joint-loss finetuning, and frozen-encoder probing, with the frozen probe best. We attribute this reversal to dropping the SSL objective during finetuning. MIAMT-MAE converges in 100 epochs, while MIAMT-JEPA requires 300. This 3× difference in finetuning time suggests

[1] http://www.charm.rhul.ac.uk/index.html

Authorized licensed use limited to: Queen Mary University of London. Downloaded on August 25,2026 at 13:55:16 UTC from IEEE Xplore. Restrictions apply.

MAE representations align more quickly with the transcription objective. We report results from the best validation checkpoint for each model. We additionally evaluate warm-start initialisation by pretraining the JEPA objective from MAE pretrained encoder weights, denoting this model MIAMT-JEPA†. All hyperparameters are kept identical to MIAMT-JEPA.

The reverse configuration (JEPA encoder weights initialising MAE pretraining) yields representational collapse ($\mathcal{R}$=79), a sharp drop in effective dimensionality, and is not investigated further. This failure likely stems from the compressed, semantically invariant nature of JEPA representations ($\mathcal{R} = 224$), lacking the fine-grained spectral detail necessary to initialise a reconstruction-based decoder, leading to immediate representational collapse.

TABLE I
LATENT SPACE EVALUATION ON NSYNTH FOR PRETRAINED (PT) AND FINETUNED (FT) ENCODERS. GEOMETRY IS MEASURED VIA ALIGNMENT $\mathcal{A}$, UNIFORMITY $\mathcal{U}$, AND PITCH LOCALITY $L_p(\Delta)$, RETRIEVAL QUALITY VIA K-NN ACCURACY κ AND REPRESENTATIONAL CAPACITY VIA RANKME $\mathcal{R}$. †INITIALISED FROM MAE PRETRAINED ENCODER WEIGHTS PRIOR TO JEPA PRETRAINING.

Model	Ckpt	Geometry				k-NN Acc. (%)		Capacity
		$\mathcal{A}_p \downarrow$	$\mathcal{U} \downarrow$	$L_p(1) \uparrow$	$L_p(12) \uparrow$	$\kappa_p \uparrow$	$\kappa_i \uparrow$	$\mathcal{R} \uparrow$
MIAMT-MAE	PT	0.197	-0.360	0.243	0.711	60.7	48.8	315
	FT	0.195	-0.372	0.240	0.704	58.3	47.9	348
MIAMT-JEPA	PT	0.541	-1.119	0.307	0.693	64.8	48.7	224
	FT	0.313	-0.610	0.371	0.753	71.9	52.9	200
MIAMT-JEPA†	PT	0.137	-0.250	0.227	0.724	56.3	51.1	282
	FT	0.189	-0.312	0.227	0.720	58.3	49.0	212

C. Objectives and Training Details

MIAMT-JEPA minimises smooth ℓ_1 (Huber) loss between predictor outputs and EMA target representations [8], [9], [11]. MIAMT-MAE optimises Mean Squared Error (MSE) reconstruction loss over masked spectrogram patches [10]. These objectives reflect different inductive biases. MIAMT-MAE focuses on pixel-space fidelity, encouraging the encoder to capture fine-grained spectral detail. MIAMT-JEPA operates entirely in representation space, bypassing signal reconstruction in favour of semantic consistency between context and target embeddings.

Both models are finetuned with an instrument-aware transcription objective [11] combining pitch activity and onset detection:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{trans}} + \alpha\, \mathcal{L}_{\text{onset}}, \quad \alpha = 0.5. \tag{1}$$

$\mathcal{L}_{\text{trans}}$ applies weighted binary cross-entropy over the pitch activity matrix $\mathbf{Y} \in \{0,1\}^{T\times N\times I}$, with instrument-activity weights $w^{\text{inst}}_{t,i} \in \{0.70, 0.30\}$ (0.70 when instrument i active, 0.30 otherwise). T, N, and I denote the number of frames, pitches, and instrument classes respectively:

$$\mathcal{L}_{\text{trans}} = \frac{1}{TNI} \sum_{t,n,i} w^{\text{inst}}_{t,i}\, \ell_{\text{BCE}}(\hat{y}_{t,n,i},\, y_{t,n,i})\,. \tag{2}$$

$\mathcal{L}_{\text{onset}}$ upweights note boundaries with a $\pm$1-frame ($\pm$10 ms) tolerance window around exact onsets to allow for minor annotation timing imprecision. Weights are $w^{\text{on}} = 0.90$ (onset) and $w^{\text{off}} = 0.10$ (non-onset) [6]:

$$\mathcal{L}_{\text{onset}} = \frac{\sum_{t,n,i} w^{\text{on/off}}_{t,n,i}\, \ell_{\text{BCE}}(\hat{y}_{t,n,i},\, y_{t,n,i})}{\sum_{t,n,i} w^{\text{on/off}}_{t,n,i} + \epsilon}, \tag{3}$$

where $\epsilon = 10^{-6}$ prevents division by zero.

D. Latent Space Evaluation

Latent representations are extracted by mean-pooling encoder outputs across the time axis followed by ℓ_2 normalisation. The NSynth test set [36] is used for evaluation, as it provides controlled monophonic recordings with ground-truth pitch and instrument labels. We report four metric categories. **(1) Representational capacity** quantifies how many effective dimensions are used. RankMe $\mathcal{R}$ [21] computes the exponential of Shannon entropy over normalised singular values (range: 1 to embedding dimension; higher indicates more dimensions actively used, less representational collapse). **(2) Geometric organisation** measures hypersphere structure via alignment $\mathcal{A}$ and uniformity $\mathcal{U}$ [24]. Alignment computes expected ℓ_2 distance between same-class pairs (range: 0-2; lower means tighter within-class clusters). Uniformity computes log-average pairwise Gaussian potential (typically -4 to 0; more negative means more evenly distributed across the hypersphere, avoiding collapsed representations). Both metrics operate on ℓ_2-normalised embeddings on the unit hypersphere. **(3) Pitch locality** $L_p(\Delta)$ measures, for each query embedding, the fraction of k-nearest neighbours in embedding space whose ground truth pitch falls within Δ semitones of the query, averaged over all queries (range: 0-1; higher means same-pitch excerpts cluster nearby). We report results for $\Delta = 1$ (semitone) and $\Delta = 12$ (octave). **(4) Retrieval quality** is measured by weighted k-NN accuracy κ [22] for pitch and instrument classification using temperature-scaled cosine similarity voting ($k = 10, T = 0.07$). Higher accuracy indicates embeddings that support better nearest-neighbour-based predictions.

E. Transcription Evaluation

Multi-instrument note-wise F1-scores are computed using `mir_eval` [37]. Frame-level predictions are post-processed using the method by Bittner et al. [6] with a threshold of 0.3, removing events shorter than 130 ms. Notes match if pitch is within 0.5 semitones, onset within 50 ms and instrument prediction is correct [2], [3]. We compare with MT3 [2], YourMT3 [3] and the MIAMT-JEPA baseline [11]. Direct comparisons should be interpreted cautiously due to differences in training data, model scale and instrument vocabularies across systems.

IV. RESULTS AND DISCUSSION

A. Learned Embeddings

Table I compares how reconstruction-based (MIAMT-MAE) and predictive (MIAMT-JEPA) objectives structure the latent space in pretrained (PT) and finetuned (FT) states. The two models exhibit complementary trade-offs across geometric organisation, retrieval quality and representational capacity.

MIAMT-JEPA achieves superior uniformity ($\mathcal{U} = -0.610$ vs. -0.372) and pitch locality ($L_p(1) = 0.371$ vs. 0.240),

Authorized licensed use limited to: Queen Mary University of London. Downloaded on August 25,2026 at 13:55:16 UTC from IEEE Xplore. Restrictions apply.

TABLE II
TRANSCRIPTION F1-SCORES. COLUMNS GROUPED BY DATASET AND INSTRUMENT.

	Maestro	*GuitarSet*	*MusicNet*		*URMP*
Model	*Piano*	*Guitar*	*Strings*	*Winds*	*Ensemble*
MIAMT-MAE	82.9	82.1	69.6	56.7	74.2
MIAMT-JEPA	81.1	82.4	69.8	60.3	72.0
MIAMT-JEPA†	85.3	84.3	74.7	65.3	78.2
SSL baseline	-	63.5	47.9	35.7	**78.9**
YourMT3	**97.0**	**91.7**	**91.3**	**83.5**	68.0
MT3	94.9	78.0	-	-	-

indicating well-spread embeddings with better pitch-space organisation. MIAMT-MAE achieves tighter within-class alignment ($\mathcal{A}_p = 0.195$ vs. 0.313), meaning same-class samples are closer together despite less uniform distribution. MIAMT-JEPA substantially outperforms on k-NN accuracy for both pitch ($\kappa_p = 71.9\%$ vs. 58.3%, $+13.6\%$) and instrument ($\kappa_i = 52.9\%$ vs. 47.9%, $+5.0\%$), confirming that superior uniformity and pitch locality translate directly to better retrieval.

MIAMT-MAE achieves higher effective dimensionality ($\mathcal{R} = 348$ vs. 200). This indicates that embeddings span a broader portion of the latent space. Surprisingly, finetuning causes the two models to diverge. MIAMT-MAE expands ($\mathcal{R}$: 315→348) while MIAMT-JEPA reduces ($\mathcal{R}$: 224→200), suggesting different adaptation mechanisms. MIAMT-MAE begins with task-relevant structure and undergoes minimal geometric change under supervision (all $|\Delta|<3\%$ except capacity $+10\%$). MIAMT-JEPA begins with loose alignment ($\mathcal{A}_p = 0.541$) and high uniformity ($\mathcal{U} = -1.119$). During finetuning it undergoes dramatic reorganisation. Alignment tightens -42%, uniformity increases $+45\%$ and k-NN accuracy improves $+7.1\%$. However, capacity reduces -11%, indicating that reconstruction pretraining produces immediately usable representations while prediction yields initially task-misaligned but adaptable ones.

MAE-pretrained initialisation of JEPA (MIAMT-JEPA†) produces intermediate pretrained capacity ($\mathcal{R} = 282$) and tighter initial alignment ($\mathcal{A}_p = 0.137$ vs. 0.541 compared to scratch JEPA), suggesting MAE initialisation preserves spectral structure that constrains early representation geometry. Finetuning reduces capacity (282→212), closely mirroring scratch JEPA's trajectory (224→200). This indicates that the JEPA objective governs adaptation dynamics regardless of initialisation.

MIAMT-JEPA performs better on 5 out of 7 latent metrics while MIAMT-MAE achieves nearly double the representational capacity. However, as shown in Section IV-B, neither property alone reliably predicts transcription performance.

B. Transcription Performance

Table II reports multi-instrument transcription F1-scores. Both MIAMT-MAE and MIAMT-JEPA substantially outperform the SSL transcription baseline [11] on three of the datasets. Despite using only 159 hours of unlabeled pretraining data, a base ViT architecture, and no dataset-specific augmentation, both models also exceed YourMT3 [3] on the URMP data (74.2% and 72.0% vs. 68.0%). Transcription results are broadly comparable between the two scratch models, with neither consistently dominating. MIAMT-JEPA outperforms on wind instruments (MusicNet Winds: 60.3 vs. 56.7, +3.6%) and narrowly on strings and guitar. MIAMT-MAE is stronger on piano (Maestro: 82.9 vs. 81.1, +1.8%) and ensemble transcription (URMP: 74.2 vs. 72.0, 2.2%). These task-specific differences suggest complementary strengths rather than a clear winner.

MIAMT-JEPA† outperforms both scratch models across all tasks, with gains of 1.9-8.6% F1 over the stronger scratch baseline per task. MIAMT-JEPA† achieves higher piano (85.3 vs. 82.9) and ensemble (78.2 vs. 74.2) scores than MIAMT-MAE despite lower capacity ($\mathcal{R} = 212$ vs. 348). This challenges the interpretation that representational capacity predicts downstream performance [21]. MAE's spectrogram-grounded structure appears directly exploitable by the JEPA objective, producing features neither objective achieves alone. These results indicate that no single latent metric reliably predicts transcription performance, and that the two SSL objectives are better understood as complementary than competing.

To verify that transcription gains stem from SSL pretraining, we additionally train encoders jointly with the transcriber from scratch, using only the transcription loss. The performance gap is largest on URMP (9.8% average decline), the smallest labelled corpus and a multi-instrument setting, and smallest on Maestro (3.2% average decline), a large single-instrument piano dataset. This indicates that SSL pretraining is most beneficial when labelled data is scarce and the task is more complex while its benefit diminishes as labelled data grows. These findings are consistent with prior work reporting that SSL is most valuable in low-resource regimes [38], [39], and that supervised-only training can approach SSL performance given sufficient labelled data [40].

V. CONCLUSION

We compared reconstruction-based (MIAMT-MAE) and predictive (MIAMT-JEPA) SSL for multi-instrument transcription. Both outperform the prior SSL baseline and exceed a state-of-the-art supervised system on URMP. Supervised-only ablations confirm these gains stem from SSL pretraining rather than the encoder architecture alone, with the benefit largest where labelled data is scarce. The SSL objectives prove complementary. JEPA is better on monophonic wind instruments, MAE on dense polyphonic content and initialising JEPA pretraining from MAE-pretrained weights (MIAMT-JEPA†) results in further consistent gains. Three insights emerge. First, no single latent metric reliably predicts transcription performance. RankMe explains differences among from-scratch models but not cross-initialisation gains. Second, MAE-pretrained initialisation provides structure directly useful for predictive pretraining. Third, the objectives exhibit opposite capacity trajectories under supervision, regardless of initialisation. Unlike the vision domain, the fact that JEPA does not dominate here suggests optimal SSL strategies are strongly modality- and task-dependent. Constrained to ViT-Base with fixed masking, our results motivate hybrid objectives and initialisation strategies that combine MAE's capacity preservation with JEPA's geometric organisation.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on August 25,2026 at 13:55:16 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 20–30, Jan. 2019.

[2] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-task multitrack music transcription," in *Proc. Int. Conf. Learn. Representations*, 2022.

[3] S. Chang, E. Benetos, H. Kirchhoff, and S. Dixon, "YourMT3: Multi-instrument music transcription with enhanced transformer architectures and cross-dataset stem augmentation," in *Proc. IEEE 34th Int. Workshop Mach. Learn. Signal Process.*, 2024, pp. 1–6.

[4] C. Hawthorne et al., "Onsets and frames: Dual-objective piano transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 50–57.

[5] Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang, "High-resolution piano transcription with pedals by regressing onset and offset times," *IEEE/ACM Trans. Audio, Speech, Lang. Process.* vol. 29, pp. 3707–3717, 2020.

[6] R. M. Bittner, J. J. Bosch, D. Rubinstein, G. Meseguer-Brocal, and S. Ewert, "A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 781–785.

[7] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1422–1430.

[8] M. Assran et al., "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15619–15629.

[9] L. Tuncay, E. Labbé, E. Benetos, and T. Pellegrini, "Audio-JEPA: Joint-embedding predictive architecture for audio representation learning," in *Proc. ICME*, 2025.

[10] P.-Y. Huang et al., "Masked autoencoders that listen," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 28708–28720.

[11] M. Pilataki, M. Mauch, and S. Dixon, "Self-supervised learning with a JEPA framework for multi-instrument music transcription," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2025, pp. 1–5.

[12] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Towards a universal audio pre-training framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2391–2406, 2024.

[13] S. Yadav, S. Theodoridis, and Z.-H. Tan, "AudioMAE : Learning better masked audio representations with swiglu FFNS," in *Proc. IEEE 35th Int. Workshop Mach. Learn. Signal Process.*, 2025, pp. 1–6.

[14] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "M2D-CLAP: Masked modeling duo meets clap for learning general-purpose audio-language representation," in *Proc. Interspeech*, 2023, pp. 1294–1298.

[15] Y. Gong et al., "Contrastive audio-visual masked autoencoder," in *Proc. Int. Conf. Learn. Representations*, 2023.

[16] A. Riou, S. Lattner, G. Hadjeres, M. Anslow, and G. Peeters, "Stem-JEPA: A joint-embedding predictive architecture for musical stem compatibility estimation," in *Proc. Int. Soc. Music Inf. Retrieval*, 2024.

[17] A. Quelennec, P. Chouteau, G. Peeters, and S. Essid, "Masked latent prediction and classification for self-supervised audio representation learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025, pp. 1–5.

[18] A. Quelennec, P. Chouteau, G. Peeters, and S. Essid, "MATPAC++: Enhanced masked latent prediction for self-supervised audio representation learning," 2025, *arXiv:2508.12709*.

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[20] J. Zhang et al., "X-ARES: A comprehensive framework for assessing audio encoder performance," in *Proc. Interspeech*, 2025, pp. 4868–4872.

[21] Q. Garrido, R. Balestriero, L. Najman, and Y. Lecun, "RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 10929–10974.

[22] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

[23] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Proc. Interspeech*, 2024, pp. 547–551.

[24] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 9929–9939.

[25] A. Kolesnikov et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[26] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[27] H.-W. Dong, C. Zhou, T. Berg-Kirkpatrick, and J. McAuley, "Deep performer: Score-to-audio music performance synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 951–955.

[28] X. Riley, D. Edwards, and S. Dixon, "High resolution guitar transcription via domain adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 1051–1055.

[29] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 45–49, doi: 10.1109/WASPAA.2019.8937170.

[30] J. Thickstun, Z. Harchaoui, and S. Kakade, "Learning features of music from scratch," in *Proc. Int. Conf. Learn. Representations*, 2017.

[31] B. Maman and A. H. Bermano, "Unaligned supervision for automatic music transcription in the wild," in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 14918–14934.

[32] Q. Xi, R. M. Bittner, J. Pauwels, X. Ye, and J. P. Bello, "GuitarSet: A dataset for guitar transcription," in *Proc. Int. Soc. Music Inf. Retrieval*, 2018, pp. 453–460.

[33] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.

[34] C. Hawthorne et al., "Enabling factorized piano music modeling and generation with the MAESTRO dataset," in *Proc. Int. Conf. Learn. Representations*, 2019.

[35] J. Hwang et al., "TorchAudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for pytorch," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2023, pp. 1–9.

[36] J. Engel et al., "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proc. 34th Int. Conf. Mach. Learn.*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., 2017, pp. 1068–1077.

[37] C. Raffel et al., "mir-eval: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retrieval*, 2014, pp. 367–372.

[38] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12449–12460.

[39] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7345–7354.

[40] K. He, R. Girshick, and P. Dollar, "Rethinking ImageNet pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4918–4927.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on August 25,2026 at 13:55:16 UTC from IEEE Xplore. Restrictions apply.