Mixing Style Representation Learning: A Stem-Aware Contrastive Framework with Structured Mix Features

Barry Cheng

Abstract

Research on "automatic mixing" remains fragmented: some works address only *single-track mastering*, while others predict effect parameters without conditioning on the pre-effect content, leaving ambiguity about what constitutes "mixing." Moreover, fully supervised studies depend on rare datasets with raw stems and detailed parameter labels, limiting scalability. This proposal introduces a **stem-aware contrastive framework** for *mixing representation learning*, aiming to learn an interpretable embedding that reflects both technical mix quality and stylistic tendencies. Unlike prior works, our model explicitly handles *separated stems* (vocals, bass, drums, other) and learns from released tracks alone—no raw multitracks—by contrasting professional mixes with synthetically degraded versions. Stage 1 performs large-scale contrastive pretraining on FMA using SCNet separations; Stage 2 freezes the encoder and regresses perceptual mix-quality scores from Mixing 101. Structured *mixing features* (dynamics, spectral balance, stereo image, inter-stem masking) are embedded and used to modulate the encoder, improving interpretability. Evaluation includes t-SNE analysis of mixing-style clustering, mix-quality prediction, human MUSHRA tests, and style-transfer demonstrations following ST-ITO.

1 Proposed Method

Stem-centric inputs. Given a stereo mixture $\mathbf{x}$, we use SCNet to separate four stems $\hat{\mathbf{s}} = \{\hat{s}^{(\text{voc})}, \hat{s}^{(\text{bass})}, \hat{s}^{(\text{dr})}, \hat{s}^{(\text{oth})}\}$. The audio encoder E_θ processes these stems individually, producing per-stem embeddings that are concatenated into a track-level representation.

Stage 1: Stem-aware contrastive pretraining on FMA. Using FMA songs, we construct degraded remixes by perturbing stems with controlled faults (e.g., gain imbalance, spectral tilt, overcompression, stereo misplacement). The encoder learns invariant representations for good vs. degraded versions via InfoNCE:

$$\mathbf{z}^{(c)} = \text{Pool}\Big(E_\theta(\hat{s}^{(c)}; \Gamma)\Big), \quad \mathbf{z} = \text{Concat}\big(\mathbf{z}^{(\text{voc})}, \mathbf{z}^{(\text{bass})}, \mathbf{z}^{(\text{dr})}, \mathbf{z}^{(\text{oth})}\big),$$

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\text{sim}(\mathbf{z}, \tilde{\mathbf{z}}^+)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{z}, \tilde{\mathbf{z}}_j)/\tau)},$$

where Γ denotes mixing-feature conditioning (Sec. 2).

Stage 2: Regression fine-tuning on Mixing 101. After contrastive training, the encoder is frozen and a regression head r_ψ predicts human mix-quality scores y:

$$\hat{y} = r_\psi(\mathbf{z}), \qquad \mathcal{L}_{\text{reg}} = \tfrac{1}{N} \sum_i (y_i - \hat{y}_i)^2.$$

This two-stage pipeline scales unsupervised learning while leveraging limited labeled data.

2 Structured Mixing Features

We compute interpretable descriptors mirroring how engineers evaluate balance and clarity. Each feature family f_j (dynamics, spectral, stereo, masking) is projected by a small MLP g_j to an embedding $\mathbf{h}_j$, and combined:

$$\mathbf{h}_{\text{mix}} = \sum_j \mathbf{h}_j, \quad (\gamma, \beta) = \text{MLP}(\mathbf{h}_{\text{mix}}),$$

which modulates the encoder through FiLM:

$$\mathbf{a}' = \gamma \odot \mathbf{a} + \beta.$$

Alternative conditioning mechanisms (prepend, cross-attention) are tested in ablation.

2.1 Loudness and Dynamics

RMS, crest factor, and loudness:

$$\text{RMS}(x) = \sqrt{\tfrac{1}{N}\sum_n x[n]^2}, \quad \text{Crest}(x) = 20\log_{10}\tfrac{\max|x[n]|}{\text{RMS}(x)},$$

$$L_{\text{int}} = -0.691 + 10\log_{10}\Big(\tfrac{1}{N}\sum_n (h_K * x[n])^2\Big),$$

where h_K is the BS.1770 K-weighting filter. Relative loudness and dynamic range are included:

$$\Delta L^{(c)} = L_{\text{int}}^{(c)} - L_{\text{int}}^{(\text{mix})}, \qquad \text{DR}_{\text{IQR}} = \text{IQR}(L_{\text{short}}(t)).$$

2.2 Spectral Balance (EQ)

For STFT magnitude $X(f,t)$ and filterbank $H_k(f)$:

$$e_k = \tfrac{1}{T}\sum_t \log(\varepsilon + \sum_f H_k(f)|X(f,t)|^2), \quad \delta e_k^{(c)} = e_k^{(c)} - e_k^{(\text{mix})}.$$

We derive low/high energy ratios, spectral tilt, and flatness per stem.

2.3 Stereo Image

$$\text{ILD} = 20\log_{10}\tfrac{\text{RMS}(L)}{\text{RMS}(R)}, \qquad \rho = \tfrac{\text{cov}(L,R)}{\sigma_L\sigma_R},$$

$$E_M = \tfrac{1}{N}\sum_n \big(\tfrac{L[n]+R[n]}{2}\big)^2, \quad E_S = \tfrac{1}{N}\sum_n \big(\tfrac{L[n]-R[n]}{2}\big)^2, \quad \text{MSR} = \tfrac{E_S}{E_M+\varepsilon}.$$

These capture spatial width, imbalance, and phase coherence.

2.4 Inter-stem Masking

To approximate how stems mask one another, define stem-wise Mel-band energies $S^{(c)}(k,t)$. For each stem i, compute dominance margin

$$\Delta_i(k,t) = S^{(i)}(k,t) - \max_{j\neq i} S^{(j)}(k,t),$$

and a logistic masking indicator

$$\mu_i(k,t) = \sigma\Big(\tfrac{\beta - \Delta_i(k,t)}{\tau}\Big),$$

weighted by activity $a_i(t)$ and perceptual band weight $A(k)$:

$$\mathcal{M}^{(i)} = \frac{\sum_{k,t} a_i(t)A(k)\mu_i(k,t)}{\sum_{k,t} a_i(t)A(k)}.$$

High $\mathcal{M}^{(i)}$ implies consistent burying of stem i by others.

3 Data

Stage 1 (contrastive). FMA dataset (44.1 kHz, diverse genres) separated by SCNet. Each track yields multiple degraded remixes via controlled perturbations: gain ± 9 dB, EQ peaking/shelf filters, compression/expansion, bandwidth limiting, stereo reverb.

Stage 2 (regression). Mixing 101 with human-rated mix-quality labels for supervised fine-tuning (encoder frozen).

4 Experiments

4.1 Evaluation Protocol

- **t-SNE clustering:** project learned embeddings to 2D to visualize genre-wise clusters. We hypothesize stylistically similar genres (e.g., rock vs. metal) exhibit nearby clusters, reflecting shared mixing traits such as dynamic range and spectral balance.
- **Mix-quality prediction:** evaluate regression correlation (R^2, Kendall τ) between predicted and human mix-quality scores on Mixing 101.
- **Mix-style transfer (ST-ITO):** adopt the Style Transfer via Iterative Target Optimization (ST-ITO) pipeline to map embeddings between source and target tracks. The resulting remixes illustrate controllable transfer of mixing style (e.g., "jazz $\rightarrow$ pop" spectral and dynamic profiles).
- **Human validation (MUSHRA):** conduct listening tests comparing original, degraded, and transferred mixes; report average MUSHRA preference scores and consistency with model predictions.

4.2 Ablation Studies

- **Feature families:** exclude one family (dynamics, spectral, stereo, masking) to evaluate its contribution.
- **Conditioning mechanisms:** FiLM vs. prepend vs. cross-attention conditioning.

4.3 Implementation Details

- Audio: stereo 44.1 kHz; STFT 1024-pt, hop 256; 128 Mel bands.
- Encoder: Transformer, 12 layers, embedding dim 768; per-stem mean pooling, concatenation.
- Feature MLPs: 2 layers, hidden size 256; FiLM head projects to (γ, β).
- Optimization: AdamW, lr 2×10^{-4}; batch 24; contrastive temperature $\tau = 0.1$.
- Regression: encoder frozen; 2-layer MLP regressor, lr 1×10^{-3}.

5 Expected Outcomes and Impact

We expect the learned space to capture both perceptual *mix quality* and stylistic dimensions observable through clustering and style-transfer. The approach mitigates data scarcity by requiring only commercial tracks and degraded augmentations, enabling scalable and interpretable mixing research. It bridges unsupervised representation learning and practical auto-mixing tools that explain and manipulate balance, tone, and space.