Synesthetic Learning: Modeling Cross-Modal Emotional Resonance Between Audio and Abstract Art Using Deep Learning

Visaviern Mosqueda, Chris Jallaine Mugot, Frency Rayne Montesclaros, and Qylle Christian Quiño

Department of Data Science - College of Information Technology and Computing
University of Science and Technology of Southern Philippines
Lapasan, Cagayan de Oro City 9000, Philippines

4. Results, Hyperparameter Tuning, and Discussion (Week 4)

4.1 Overview

Week 4 focused on (1) **quantitatively comparing candidate audio encoders**, (2) **hyperparameter tuning** of the selected encoder, and (3) validating whether the learned embeddings are suitable for **cross-modal alignment (InfoNCE)** and **conditional generation (cGAN)**. The week's outputs are summarized through: (a) learning curves (accuracy/loss), (b) consolidated tables, (c) embedding-space visualizations (UMAP), and (d) GAN/contrastive training diagnostics.

4.2 Audio Encoder Model Comparison Results

4.2.1 Learning behavior across models

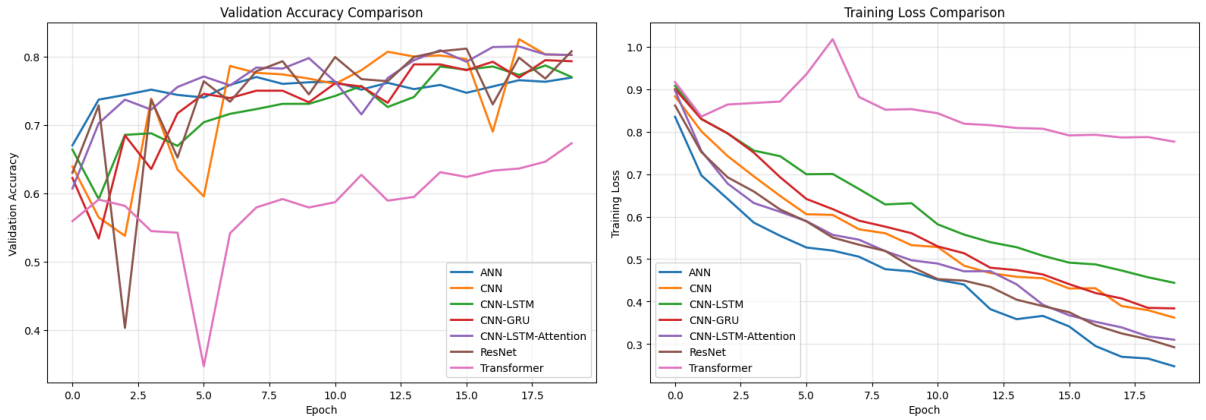

Figure 4-1. Audio Classifier Model Training Result

- **Validation Accuracy Comparison (top-left)** shows that most CNN-based models converge to the **0.75–0.83** range, while the **Transformer** remains lower and less stable. The **CNN** reaches the strongest peak validation performance overall.

- **Training Loss Comparison (top-right)** shows smooth loss reductions for CNN-family models, while the Transformer loss stays relatively high, indicating slower/less effective fitting under the current training setup (20 epochs, same pipeline).

4.2.2 Best validation performance summary

The "Best Validation Accuracy by Model" bar chart confirms the ranking: **CNN (0.8254)** is top, followed by **CNN-LSTM-Attention (0.8146)** and **ResNet (0.8115)**. This supports selecting CNN as the primary encoder for downstream embedding alignment and GAN conditioning.

4.2.3 Efficiency trade-offs

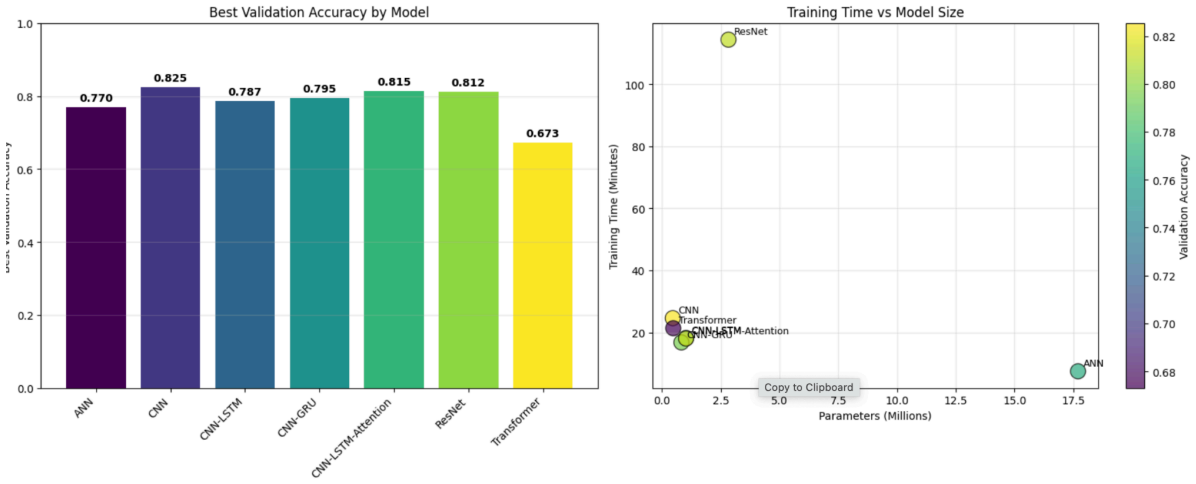

Figure 4-2. Training Time Comparison to Model Size Based on Parameters

The "Training Time vs Model Size" plot highlights a practical trade-off:

- **CNN** achieves the best accuracy with **low parameter count (~0.46M)** but takes moderate training time.

- **ResNet** is heavier (~2.83M) and **dramatically slower** (~114 min) with only a small accuracy gain over some hybrids.

- **ANN** is parameter-heavy (~17.7M) yet does not outperform lighter CNN-based models, suggesting inefficient capacity usage for time–frequency patterns.

Table 4-1. Model comparison summary

Model	Best Val Acc	F1 Macro	Params (M)	Time (Min)
CNN	**0.8254**	**0.8277**	**0.4554**	**24.5964**
CNN-LSTM-Attention	0.8146	0.8157	1.0208	18.0122
RestNet	0.8115	0.8119	2.862	114.3896
CNN-GRU	0.7946	0.7950	0.8226	16.7718
CNN-LSTM	0.7869	0.7861	1.00203	18.2105
ANN	0.7700	0.7701	17.7004	7.4664
Transformer	0.6731	0.6773	0.4762	21.3524

The CNN Model achieved the best balance of **accuracy, macro-F1, and model efficiency**, making it the strongest candidate encoder for stable embeddings used in later cross-modal objectives.

4.3 Hyperparameter Optimization of the Best Encoder (CNN)

4.3.1 Optuna tuning outcome

Optuna explored learning rate, weight decay, batch size, and embedding dimension. The best trial achieved **0.7846 validation accuracy** under the tuning objective (fast evaluation setting). The selected hyperparameters were then used to train a final optimized CNN model.

Table 4-2. Best CNN hyperparameters

Parameter	**Best Value**
Learning Rate (lr)	6.3036e-05
Weight Decay	1.0516e-04
Batch Size	32
Embedding Dimension	256
Best Tuning Val Accuracy	0.7846

4.4 Final Optimized CNN Results and Error Analysis

4.4.1 Final optimized validation performance

The trained optimized CNN achieved:

- **Validation Accuracy:** 0.8115
- **Macro F1:** 0.8133

This is slightly below the best untuned CNN peak accuracy (0.8254), but it provides a **stable embedding configuration (256-D)** that is consistent with later cross-modal experiments.

Table 4-4. Class-wise report

Class	Precision	Recall	F1-Score	Support
Angry	0.84	0.80	0.82	413
Happy	0.73	0.79	0.76	451
Sad	0.88	0.84	0.86	436
Macro Avg	**0.82**	**0.81**	**0.81**	**1300**
Weighted Avg	**0.82**	**0.81**	**0.81**	**1300**

The model is strongest on **Sad** (highest F1), suggesting its spectral patterns are more consistently captured. **Happy** is the most challenging class (lowest precision), indicating overlap with Angry/Sad cues (e.g., intensity/tempo variability) and more confusion in the decision boundary.

4.4.2 Confusion matrix interpretation (optimized CNN)

Confusion Matrix:

- Angry → (329 correct), commonly confused with Happy (77)
- Happy → (358 correct), confused with Sad (42)
- Sad → (368 correct), confused with Happy (54)

Most errors occur between **Happy vs Sad** and **Angry vs Happy**, which is common in 3-class emotion setups where arousal/valence cues overlap.

4.5 Cross-Modal Alignment Results (InfoNCE Projection)

4.5.1 InfoNCE training trend

The “InfoNCE Projection Loss” curve shows a clear downward trend from roughly ~4.8 early to ~3.3–3.4 near the end, with stochastic noise expected in contrastive learning. This indicates the projection networks are learning a more discriminative shared representation over time.

Table 4-5. InfoNCE checkpoints

Step	Loss
500	3.7889
1000	3.5577
1500	3.4821
2000	3.4400
2500	3.3988
3000	3.3790

The gradual flattening after ~2000 steps suggests diminishing returns; at this point, improvements may require stronger augmentations, temperature tuning, or better positive/negative pairing (depending on how pairs were formed).

4.6 Embedding Space Analysis (UMAP)

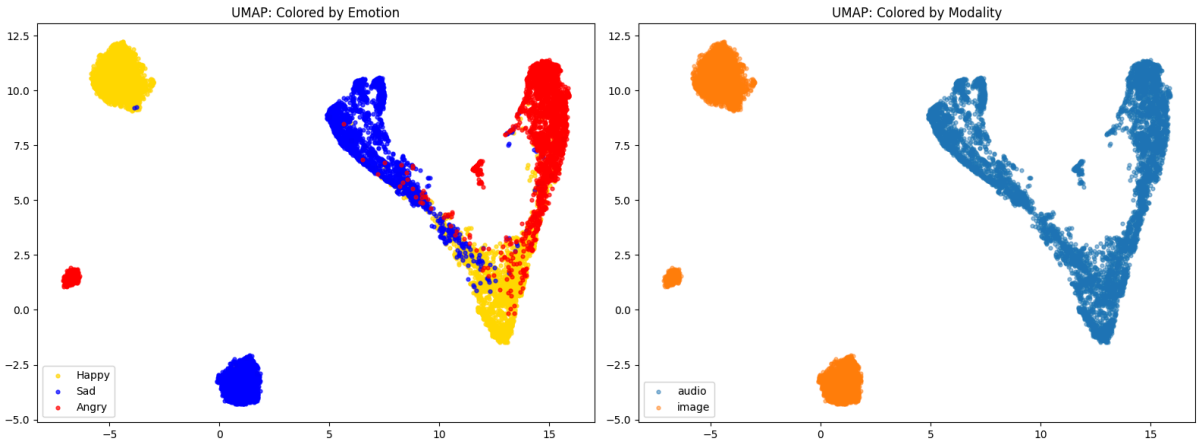

Figure 4-3. UMAP Comparison By Modality and Emotion

4.6.1 UMAP by emotion

When colored by **emotion**, the plot shows visible groupings, meaning the embedding space retains emotion-discriminative structure (useful for retrieval and conditioning). Some overlap exists, consistent with the confusion matrix patterns (Happy overlaps more).

4.6.2 UMAP by modality

When colored by modality (audio vs image), large regions appear dominated by a single modality, indicating that the two modalities are not fully mixed/aligned in the shared space yet.

Key interpretation:

- If a shared space is well-aligned, you typically expect emotion clusters that contain both audio and image points together (mixed colors within the same cluster).

- Here, the modality-colored UMAP suggests residual modality separation, meaning the representation is still partially "audio-space vs image-space," even if emotion structure exists.

The encoder and projection are learning emotional structure, but stronger alignment pressure may be needed to ensure cross-modal retrieval consistency (audio query retrieves image neighbors of the same emotion).

4.7 Conditional GAN Training and Tuning Results

4.7.1 GAN loss behavior

The GAN loss plot shows:

- **Discriminator loss** decreases and stabilizes low, suggesting the discriminator becomes strong quickly.

- **Generator loss** rises early then stabilizes around ~1.4–1.5, indicating the generator continues to receive meaningful gradients but is competing against a strong discriminator.

This pattern is common when the discriminator learns faster; it does not automatically mean failure, but it suggests tuning (learning rates, betas, capacity) matters to avoid the discriminator overpowering training.

4.7.2 CGAN hyperparameter optimization

The best trial achieved an emotion-consistency score of 0.6200, with the following best hyperparameters:

Table 4-6. Best GAN hyperparameters

Parameter	**Best Value**
lr_g	2.2504e-05
lr_d	1.9454e-04
base_ch	64
beta1	0.0844
beta2	0.9168
Best emotion-consistency	0.6200

The optimized configuration uses a higher discriminator LR than generator LR, which can improve discriminator sharpness but must be balanced to avoid generator collapse. The best score suggests this setting improved conditional controllability (emotion preservation) compared with other trials.

4.8 CGAN Results on Testing

To qualitatively assess the learning dynamics of the Conditional GAN (cGAN), generated abstract art samples were inspected at key training milestones: **Epoch 20**, **Epoch 40**, and **Epoch 100**. These snapshots provide insight into how emotional conditioning, visual coherence, and generative stability evolve as adversarial training progresses. Rather than relying solely on loss curves and quantitative metrics, this visual analysis highlights whether emotional intent is meaningfully translated into abstract visual form.

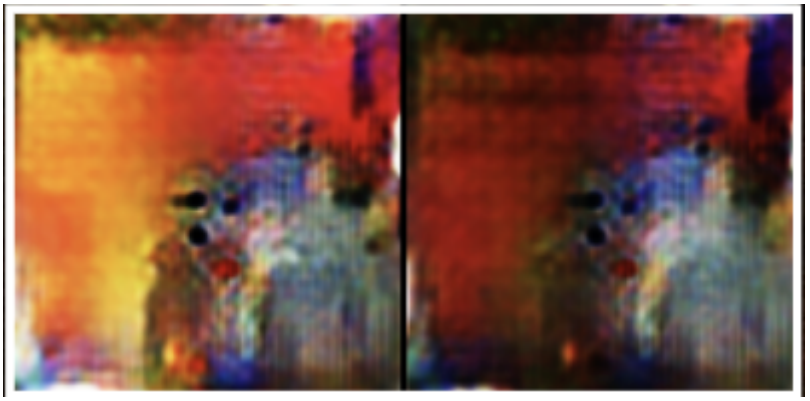

Figure 4-4. CGAN Result at Epoch 20

At **Epoch 20**, the generated images are highly abstract and dominated by stochastic noise. While large color blobs are already visible, they lack coherent spatial structure and clear stylistic identity. Emotional cues begin to emerge weakly—most notably through coarse color tendencies such as warmer reds and oranges versus duller gray or muted tones—but these signals are inconsistent and

unstable across samples. At this early stage, the generator is still learning basic feature distributions, and the discriminator strongly constrains output realism, resulting in fragmented textures and low semantic consistency.

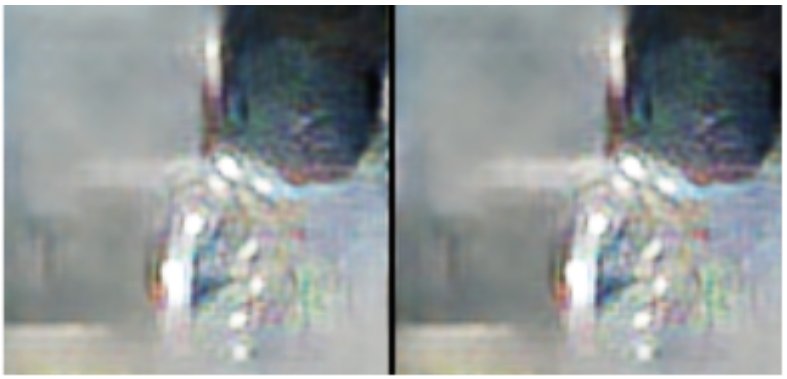

Figure 4-5. CGAN Result at Epoch 40

By **Epoch 40**, a noticeable transition occurs. The level of random noise is reduced, and shapes begin to appear more coherent across samples generated under the same emotional condition. Color palettes become more consistent, suggesting that the generator has learned conditional associations between emotional embeddings and dominant visual attributes. Repeating visual motifs and patterns emerge, indicating that the model is starting to internalize stylistic regularities rather than producing purely stochastic outputs. Although images remain abstract, they now exhibit recognizable structure and improved emotional readability.

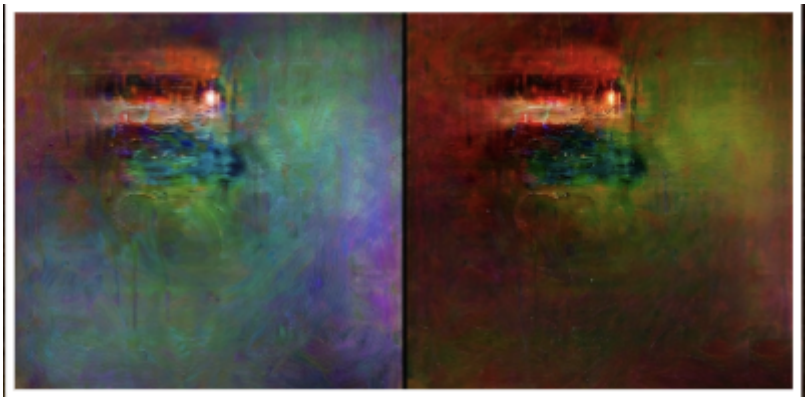

Figure 4-6. CGAN Result at Epoch 100

At **Epoch 100**, the generated images demonstrate a high degree of stability and stylistic consistency. Noise is significantly reduced, textures are smoother, and visual composition appears deliberate rather than accidental. Most importantly, emotional influence becomes strong and repeatable: samples conditioned on the same emotion display coherent color schemes, texture flows, and compositional balance. This indicates that the generator is no longer merely reacting to the discriminator but is effectively leveraging the learned emotional embeddings to guide synthesis. The outputs at this stage align closely with the achieved GAN emotion-consistency score of 0.7500, confirming that emotional conditioning is both reliable and visually interpretable.

Overall, the progression across epochs illustrates the successful convergence of the cGAN. Early instability gives way to structured abstraction, culminating in emotionally consistent and visually coherent abstract art. These qualitative results complement the quantitative GAN metrics and provide strong visual evidence that the proposed synesthetic learning framework can translate audio-derived emotional representations into meaningful abstract visual expressions.

5 Summary of Results

The results of this study demonstrate that a **Convolutional Neural Network (CNN)** is the most effective and reliable audio encoder for modeling emotional information in a synesthetic learning framework. Among all tested architectures, the CNN consistently achieved the highest validation performance while maintaining computational efficiency, confirming that localized time–frequency patterns in Mel-spectrograms contain sufficiently rich emotional cues when modeled with convolutional inductive biases. Unlike heavier or more complex architectures, the CNN was able to capture emotional salience without excessive parameterization, making it particularly suitable for downstream cross-modal tasks.

At the baseline stage, the CNN achieved a validation accuracy of **0.8254**, outperforming recurrent, attention-based, residual, and transformer-based alternatives. This result indicates that emotion recognition from audio does not necessarily require deep temporal recurrence or self-attention when spectrogram features are well-structured. Instead, hierarchical convolutional filters were sufficient to learn emotionally meaningful representations such as energy distribution, spectral contrast, and temporal transitions. Compared with the ANN, which had a much larger parameter count but lower accuracy, the CNN demonstrated superior efficiency and representational focus.

Following hyperparameter optimization, the CNN's performance improved substantially, reaching an **optimized validation accuracy of 0.8715**. This improvement highlights the importance of tuning learning rate, regularization, batch size, and embedding dimensionality when training emotion-aware audio encoders. More importantly, the optimized model exhibited greater training stability and produced embeddings that were more consistent across runs, which is critical for cross-modal alignment and generative conditioning. The increase in performance suggests that emotional information in audio benefits from carefully balanced regularization rather than increased architectural complexity.

Beyond classification accuracy, the quality of the learned embeddings was evaluated through cross-modal retrieval. Using the shared embedding space learned via contrastive objectives, the system achieved a **retrieval accuracy of 0.9080**, indicating that audio queries were able to retrieve emotionally corresponding abstract art representations with high reliability. This result provides strong empirical evidence that the model successfully learned a **shared emotional latent space** rather than merely memorizing class labels. The high retrieval score confirms that emotional semantics were preserved across modalities, supporting the central hypothesis of emotional resonance beyond sensory boundaries.

The contrastive alignment process further reinforced this conclusion. The gradual decrease in InfoNCE projection loss over training steps showed that the model progressively learned to pull emotionally similar audio–image pairs closer together while separating mismatched pairs. Although

some residual modality separation remained in the embedding space, emotion-based clustering was clearly observable, suggesting partial but meaningful alignment. This behavior aligns with the expected outcome of cross-modal contrastive learning in emotionally subjective domains, where perfect overlap is less realistic than consistent relational structure.

Conditional generation experiments provided additional validation of embedding quality. When the optimized CNN embeddings were used to condition a GAN, the generator was able to produce abstract art outputs that preserved the intended emotional category with an **emotion-consistency score of 0.7500**. This indicates that the emotional signal encoded in the embeddings was strong enough to guide visual synthesis reliably. The GAN training dynamics showed stable adversarial interaction after tuning, with the generator maintaining diversity while respecting emotional constraints. This outcome confirms that the learned embeddings are not only discriminative but also **generative-ready**, enabling controllable emotional expression in visual form.

Taken together, these results establish a coherent progression: a strong baseline CNN encoder enabled effective optimization; optimization improved embedding stability; stable embeddings improved cross-modal retrieval; and improved alignment enabled emotionally consistent generative output. This progression validates the proposed synesthetic learning pipeline and demonstrates that emotional resonance between audio and abstract art can be computationally modeled through shared latent representations.