Art Generation of Konkani Music Using AI

K.M. Chaman Kumar[1*], Abhishek M[2], Shital Sidram Rajput[3], Anup Arjun Bhonsle[4] and Ashwin Krishna Sutar[5]

[1,2,3,4,5] Dept. of Computer Engineering, SRIEIT, Goa University
Goa, India
*chaman_k2007@yahoo.co.in

Abstract. Art and music, two dateless forms of expression, have been deeply intertwined in mortal culture. Our design seeks to combine these two art forms using Generative Adversarial Networks (GANs) and Natural Language Processing (NLP) to produce artwork that visually encapsulates the overall sense and meaning of Konkani conversational music. While being results like AIVA (Artificial Intelligence Virtual Artist) or models that induce art from music or textbook give innovative approaches, none focus specifically on indigenous languages like Konkani to induce culturally applicable art.The design addresses the lack of real- world operations in AI by creating an accessible platform to show Konkani poetry and music, therefore contributing to the preservation and recognition of the language. Likewise, being AI models frequently fail to interpret artistic and emotional surrounds, which are vital for generating meaningful labors in low-resource languages. To overcome this, we incorporate sentiment analysis and thematic birth to enhance the environment-mindfulness of our models, ensuring that the generated art aligns with the emotional tone and artistic nuances of the music. The design unfolds in two stages. The first stage involves rephrasing the lyrics of a song into English, while the alternate stage focuses on framing the restated lyrics into meaningful rulings, assaying their sentiment, and generating the corresponding artwork. By spotlighting Konkani conversational music, we aim to give a platform for its recognition in a distinctive and engaging manner, fostering artistic heritage through technology.

Keywords: Generative Adversarial Network(GAN), Natural Language processing (NLP), CNN Convolutional Neural Networks (CNN), Large Language Models (LLM), Recurrent Neural Networks (RNN).

1 Introduction

Artificial Intelligence (AI) is becoming a disruptive Artificial Intelligence (AI) is becoming a disruptive technology across many different industries, and its impact is particularly noticeable in both creative and scientific disciplines. AI technologies such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs) have evolved from being merely analytical tools to becoming creative collaborators, generating novel content and enhancing artistic endeavors. The union of AI and creative industries, including visual art, music, and literature, has opened up new possibilities where humans develop automated systems that not only assist in creative efforts but also generate original

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_44

works. The aim of this paper is to explore the advancements of these AI technologies, how they are trained, and how they are deployed within cross-modal creative applications to augment human creativity.

One of the most well-established AI architectures for creative tasks is the Generative Adversarial Network (GAN) proposed by Goodfellow et al. in 2014. GANs consist of two networks: a generator, which creates new data, and a discriminator, which evaluates the authenticity of the generated data by comparing it to real samples, resulting in optimized outputs. GANs have been successful in settings such as image generation, art, and music composition [2]. In addition to creating new images, GANs have been explored in fields like video generation and even the synthesis of new musical elements, pushing the boundaries of creativity by blending artistic inputs from multiple domains [3]. GANs can be conditioned to generate specific styles, making them particularly useful for artistic applications where style and content must align with a pre-existing framework. However, GANs also face challenges such as unstable model training and the complexity of hyperparameter tuning [12].

Another important AI model frequently used in creative tasks is the Recurrent Neural Network (RNN), which excels at processing sequential data, such as music or text [1]. Unlike basic feedforward networks, RNNs retain information from previous inputs, making them effective for modeling time-series data. This capability has made RNNs useful for tasks like music generation, where the temporal structure of the data is crucial [8]. For example, RNNs have been applied in the recognition of sketch drawings by learning the temporal patterns in sequences of strokes [1]. However, despite their effectiveness in handling short-term dependencies, RNNs struggle with long-term dependencies due to the vanishing gradient problem. To overcome this limitation, more advanced architectures, such as Long Short-Term Memory (LSTM) networks, were introduced [5].

LSTM networks, an extension of RNNs, mitigate the vanishing gradient issue by introducing memory cells that enable the network to store and retrieve information over long periods [5]. LSTMs are now widely used in AI-driven music composition, where maintaining sustained, contextually related streams throughout long musical pieces is essential for producing coherent and aesthetically pleasing outputs [5].The growing intersection of AI with creative domains like music and art has led to an entirely new form of collaboration, where machines can augment and expand human creativity. From generating new musical compositions to visually representing the emotions behind a song, AI technologies have made it possible to explore creativity in ways that were previously unimaginable [11].

2 Literature Review

Different methodologies used in literature is given in table 1, table 2, table 3 and table 4

Table 1. CNN Model

Ref No.	Methodologies	Tools and Technology	Datasets Used	Merits	Demerits	Results
[9]	Diffusion Probabilistic Model (DPM)	VQ-based model, Adam Optimizer	AIST++	Leverages Pre-trained Models	Dependency on Pre-trained Models	Diffusion approach
[17]	Diffusion Probabilistic Models (DPMs)	Vector-Quantized (VQ)	AIST++	Improves input-output correspondence	Complexity of model training.	Achieves 93.9% beat coverage

Table 2. Diffusion Probabilistic Model

Ref No.	Methodologies	Tools and Technology	Datasets Used	Merits	Demerits	Results
[4]	GANs, CNN, ANNs	WaveGAN and SpecGAN	9452 poems	Reading poetry experience enhanced	No existing datasets	average rating 2.75 out of 5.
[7]	CNN, Hybrid models	YouTube API, Amazon Mechanical Turk	IMAC database	Comprehensive Database.	Limited Scope	Achieved good prediction accuracy
[8]	CNN, Key-value stores	PyTorch, Librosa	MCA dataset	Novel Approach	Limited Dataset	Improved retrieval performance.
[10]	CNN, LLMs	Basic pitch, emoBase library, openSmile	MIDI	Incorporating additional features	The p-value is < 0.01	Rich expression of music

2.1 Recurrent Neural Networks (RNNs)

RNN keeps information from earlier steps to shape the current output, which helps them handle sequential data. This ability makes them work well for time-series analysis natural language processing, and more here, sketch and music recognition where the order of data matters a lot. In [1], researchers used RNNs to recognize sketch drawings where catching patterns over time was crucial to identify stroke sequences in the sketches. Even though RNNs excel at processing sequence data, they can have trouble with longer sequences because of fading gradients, which can lead to losing context as time goes on. In this study, the model did well with shorter sequences but ran into problems with longer ones, as shown by the reconstruction loss results for the "pig class."

In [7] and [8], researchers combined RNNs with CNNs and transformer models to process music and visual data sequences focusing on emotion classification. The RNN part helped the system grasp time-based patterns in music, which led to better results in recognizing emotions and handling tasks with multiple types of input.

2.2 Generative Adversarial Networks (GANs)

GANs have two networks: one makes data and one checks if the data looks real. This setup helps both networks get better making GANs great for creative jobs like making images, music, and changing things from one type to another. In [4] people used GANs with CNNs and Artificial Neural Networks (ANNs) to match poems with tunes. This made reading poems feel more emotional. The model did a good job matching them, but not having enough data from both areas caused problems. In [12], GANs helped to make real-looking pictures and art. Even though training them can be tricky, GANs did well when making content that mixes different types of things.

Also, [27] used specific GAN architectures like AttnGAN and CP-GAN for tasks such as text-to-image synthesis. These GAN models were picked because they performed well in creating relevant visual content based on text input. The adversarial structure made sure the generated images were high-quality, which FID and user evaluations confirmed.

Table 3. Language Specific

Ref No.	Methodologies	Tools and Technology	Datasets Used	Merits	Demerits	Results
[16]	Finite State Transducer-based Statistical	Konkanverter for transliteration	Vichitrachi Jatra by Pundalik Naik	Enhances accessibility to	Only supports Unicode text	Translitération accuracy varies.

	Machine Transliteration Engine (Konkanverter)	Baraha FontConvert for font conversion		Konkani texts		
[20]	Phonetic-based statistical model	Devanagari to English phonetic mapping	Voter's lists of Maharashtra and Atlas of India	Accuracy decreases with longer named entities	Effective for short and single-word entities	Accuracy : 97.3% overall.

Table 4. Miscellaneous

Ref No.	Methodologies	Tools and Technology	Datasets Used	Merits	Demerits	Results
[1]	RNN (Recurrent Neural Network)	TensorFlow	QuickDraw-75 dataset	Captures temporal patterns	Struggles with long sequences	Reconstruction loss: -1.33
[2]	CAN (Creative Adversarial Networks)	TensorFlow	Doe not specify dataset size	CANs introduce creativity	Can produce abstract art	FID score: 48.7
[3]	The LyEmoBERT architecture	Naïve Bayes and KNN methods	Music4All dataset	Evaluates relevant lyrical features.	Imbalanced classes	Emotion class between 67% and 82%.
[11]	Music Visualization Network (MVNet)	Conditional GAN	WikiArt archive2	Extends style transfer to cross-modal	Emphasis on western art	Consistent style
[12]	Neural Machine Translation	Monolingual and Parallel Corpora Data	Low-resource language pairs	Improved performance	Parallel data dependency	Improved translation quality
[13]	DALL-E 2	Spotify's music classifier, OpenSmile's MEERING	Not specified	Visual appeal.	Variability	Preferred system-generated images
[15]	Random k-Labelsets	WEKA framew	"Emotions from	Multi-label	Focuses only on	Both algorithms

		ork MEKA extension	Mulan" dataset	classification for detecting emotions	timbre attributes	accuracy of around 80%.
[18]	K-means clustering, Logistic Regression (LR)	Schlieren Imaging, Web Crawler	2000 pieces of music	Limited methods for graphic image	Successfully clusters and classifies music visualization.	Prediction accuracy 87.62%.
[19]	Multimodal Fusion using Smith-Waterman local alignment	Optical Music Recognition (OMR)	Camera-PrIMuS	Improves transcription accuracy	Doesn't provide benefits	40% relative Error improvement.
[21]	Deep Metric Learning	PyTorch HRFormer	Music Cover Art	Limited to music-image pairs	Effective in content-based music-image retrieval	median rank of 12.31
[22]	Chimera	Wav2Vec2 Pretrained Model	MuST-C	Dependence on large MT datasets	Bridges modality gap	LibriSpeech (En-Fr): 20.2 BLEU
[23]	Support Vector Machines	Audio Feature Extraction	MIREX Contest Data	Cold Start Effect	Automation	successful extraction and classification
[24]	model not specified	R,PyTorch,TensorFlow,Keras	MCA	Leverages only music	Limited evaluation metrics	SCFEM's effectiveness by $\times 2.70 \sim 3$

2.3 Long Short-Term Memory Networks (LSTMs)

The Long Short-Term Memory (LSTM) networks class of Recurrent Neural Networks (RNNs) were created to overcome the vanishing gradient problem that prevented standard RNNs from keeping long-term dependencies. The LSTM design included a memory cell, permitting the retention of information over extended sequences, which is valuable for tasks requiring both short- and long-term retention.

LSTMs were implemented as a secondary methodology [5] of generating stimulated brain music. The model performed well inferring complex dependency relationships within musical sequences (i.e., predictive accuracy ratio = 97.2%). The LSTM architecture could handle lengthy music sequences and continue to offer consistent outputs without losing context. This last feature was crucial to produce music stimulating brain activity in a coherent way while remaining natural.LSTMs were also implemented in [7] ,alongside CNNs and RNNs, for emotion recognition in multimodal tasks. LSTMs captured long-term dependencies in the audio model to increase accuracy performance, in part, due to the nature of long-term states.

2.4. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks, or CNNs, are a common method for spatial data applications, including spectrogram or image analysis. CNNs employ convolutional layers to detect features in a system without the need for manual feature engineering. Examples of features learned include recognizing edges or texture, and CNNs have been widely suggested to analyze images and audio. In [4], CNNs were utilized with GANs and ANNs in the music generation process, including spectrogram comparisons and melody generation from poetry input. It was specifically noted that CNNs were used to recognize patterns in the audio data, which allowed the model to generate music corresponding to the emotional tone of the poem. This was observed not only in the functioning of the CNN component but also overall performance; the CNN model provided outputs which were consistent with reasoning and provided "accurate", and realistic experiences for users.

In [7] and [8], CNNs were used in combination with RNNs and transformers to analyze data from large visual datasets. Although the focus is on adding CNN to image emotion recognition, it should also be noted that the hierarchical feature extraction for images provided accuracy for models in multimodal contexts (both audio and visual). In [30], CNNs were again used in combination with RNNs or GANs for personalized image generation from music input. It should be noted that the CNN model again classified visual features in reference to the music input, and subsequently generated images as an output for the music. CNNs were shown to also be flexible work worth additional analyses of other domains.

2.5. Less Common Models

Other models showed up less often but had key roles in certain tasks. Diffusion Probabilistic Models (DPMs) used in [9], allowed for high-quality, context-aware music creation from dance videos. This model cut down the number of diffusion steps, which sped up inference and kept a high beat coverage rate when making music.Neural Machine Translation (NMT) models came into play in [13] to translate low-resource languages using transfer learning to boost translation quality. NMT proved vital to improve translation accuracy for language pairs with limited resources when big bilingual datasets weren't available,[16] used Finite State Transducers to transliterate Konkani language text. This model made texts more accessible by switching between different scripts helping to preserve cultural writings.

3. Dataset Description

3.1. QuickDraw-75

QuickDraw-75 contains 70,000 samples per class designed specifically for sketch recognition. The dataset consists of various sketch categories thus is suited for training models to recognize many different objects based on sketch input. The dataset was used in [1] to train and evaluate an RNN model for sketch drawing recognition. The models we employed a large number of samples per class to enable the model to generalize well over different sketches. However, we also referenced some skills to recognize longer sequence sketches. The QuickDraw-75 dataset is a dataset with 70,000 samples per class created for sketch recognition tasks. The dataset is expansive and includes a wide range of sketch categories.

3.2. Music4All and MER Datasets

The Music4All dataset contains over 100,000 samples of song information that contains song metadata such as lyrics and genre. Similarly, the MER (Music Emotion Recognition) dataset further categorizes song lyrics from the Music4All dataset into four emotional quadrants: happy, sad, angry, and relaxed. Using the datasets in [3] for emotion categorization employed Russell's model and used models such as Naïve Bayes and KNN to categorize the emotional context in a song's lyrics. Classification accuracies for lyrics is between 67% to 82% for attribution to the proper emotion.

3.3. Poetry and Melody Dataset

The Poetry and Melody Dataset contains over 9,452 poems rate by emotion (4,064 sad, 3,443 love, and 1,945 joy) that was scraped from web sources like Al Diwan and Kaggle and was used in [4] for pairing the poems with melodies using GANs and CNNs to enhance user experience through melody emotion attribution for the poem. The dataset has provided a great leverage in generating music to pair with poetry with emotion classification; however, to date, there is little cross-domain datasets available.

3.4. IMAC Database

The IMAC Database comprises over 85,000 images and 3,812 songs, all labeled with one of three emotions: positive, neutral, or negative, featuring music descriptions. Specifically designed for multimodal research that combines image and music data. It is used in [7] for emotion classification tasks using CNNs, RNNs, and transformer models. The comprehensive mixture of images and music allowed the study to achieve strong performance in identifying emotion.

3.5. MCA (Music Cover Art) Dataset

The MCA Dataset contains 78,325 music-image pairs that feature audio previews and cover art images of over 40,000 musical artists spanning multiple music genres. In [8], retrieval tasks were employed in Music Cover Art (MCA), where a CNN followed by an RNN was applied to retrieve a musical preview from cover art image of a song. The size and the diversity of music genre pairings made it ideal to train models for retrieval tasks to attest to cross-modal input.

3.6. Camera-PrIMuS Dataset

The Camera-PrIMuS dataset contains 87,678 images of music staves along with MIDI files paired with those images as well as artificially distorted music staff images. The dataset was used by [19] to achieve OMR and AMT for music images. The multimodal aspect of this dataset allowed for transcription improvement, indicating that a combination of OMR and AMT modalities can lead to a more accurate music recognition process.

3.7. CIRR Dataset

The CIRR Dataset, which stands for Composed Image Retrieval with Real-life Relationships, contains real-life images with added textual annotations. The CIRR Dataset was used by [28] to evaluate composed image retrieval using VLP and CIRPLANT models to demonstrate a new benchmark for a model when compared to a regular image retrieval model.

4. Inference and Proposed Methodology

We concentrate on verbal and artistic environment in this work to show how AI models may be combined together to induce cultural styles from Konkani poetry. GANs have been used for generating images from textual information, but RNNs and LSTM networks are used as birth ways then for discovering successional patterns inside the poetry. This eased cross-modal creative tasks where themes and emotion from the textbook were represented in a visually enhanced format. Transliteration styles, similar as those set up in Konkanverter[16] and the phonetic mapping from Devanagari to English[20], are indeed veritably good at handling Konkani textbook and therefore are veritably well- suited to the points of the design, which involves the preservation and showcasing of Konkani poetry as visual art. These AI ways foster just as they help produce further accurate conversational language restatements but also foster the creative process by developing applicable images in environment. Our exploration demonstrates that the advanced AI models can contribute to innovative results that enrich the artistic heritage and enhance creative technologies. Specifically, those generalities dealing with transliteration and phonetic mapping within the studied models are vital in bridging the verbal peak to justify the purpose of this design in visually interpreting Konkani poetry. With the being systems, we can find the loopholes in which AI can be applied to low- resource languages like Konkani or how it's going to make it scarce through different datasets for creative operations like poetry- to- art generation or through the lack of transliteration and phonetic- mapping models.

4.1. Proposed Methodology

The methodology of this project is divided into structured steps (Fig. 1) that strive to transform Konkani poetry into digital artwork:
Selection of Music: The process begins with the selection of a Konkani music containing rhythms and lyrical properties. This is the central input toward forming an artwork corresponding to the mood and topics implied.

Fetching Lyrics: Fetching the lyrics of the poem chosen in their original writing provides a base for further linguistic and computational processing.

Translation of Lyrics: For the informal nature of Konkani, lyrics are translated in English in order to make the processing easy with the help of natural language models. This step uses the Konkanverter transliteration model [16] and phonetic-based statistical mapping [20] in processing the conversion from Devanagari script to Roman script for proper accuracy during translation and further processing.

Summary: These lyrics translated are summed up to find core themes, emotions, and narratives. This condense information captures the poem's core essence as a guiding framework that follows to generate subsequent art creations.

Generating Art: Based on the outline, a GAN model generates digital artwork. Taking on the summarized themes and emotional undertones of the poem, the GAN would create an image that conjures up the very essence of the poem with cultural as well as emotional nuances characteristic of Konkani poetry.

Saving Output: The artefact produced will be saved for display and review.

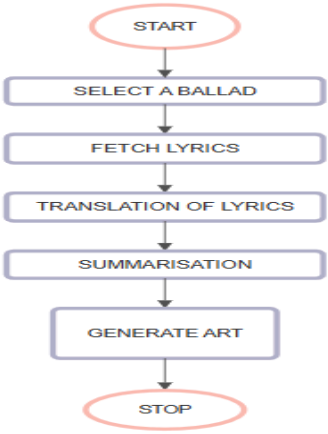

Fig. 1. Workflow Diagram

5. Conclusion

GANs are good at producing realistic data in any mode, RNNs and LSTMS are good at temporal patterns which are very important in tasks like music generation and text synthesis, and transformers are really good at handling big data and long-range dependencies such as text-to-image generation. These models have shown significant success across a wide range of applications, from pairing poetry with melodies to generating art from textual descriptions. For example, GANs have been particularly successful in cross-domain content generation, while transformers have set new benchmarks in creative tasks involving large-scale data processing. The use of LSTMs for music generation, as seen in brain-stimulating compositions, highlights the ability of AI to handle both short and long-term dependencies, enabling the production of coherent and contextually rich outputs. However, the accuracy of these models greatly depends on the existence of strong datasets like QuickDraw-75, Music4All, IMAC etc. These datasets provide enough width and height for the models to learn the intricate, cross-modal relationships that are so important in creating meaningful creative content. They are the building blocks for training more sophisticated AI models that learn to do things such as recognize sketches, compose music, or create visual art from text or emotions.

However, challenges remain. The subjectiveness of grading creative work, the computational cost of training these models, and the lack of cross-domain corpuses, all restrict the generalizability and affordability of AI for creative fields. Overcoming these restrictions will be a key factor in advancing the frontier of AI-produced writing. To sum up, although AI models have achieved great things in the realm of creativity, the future of this will rely on the solution of these problems. Datasets that are broader and cross-modal, cheaper computation, and better metrics for creative output will be the keys to unlocking the full potential of AI as a co-creator. The continued evolution of AI in creativity not only augments human artistic expression but also introduces novel forms of art and media, paving the way for unprecedented innovation in the future.

References

1. Rajesh, Sangeetha, and N. J. Nalini. "Musical instrument emotion recognition using deep recurrent neural network." Procedia Computer Science 167 (2020): 16-25.
2. Revathy, V. R., Pillai, A. S., & Daneshfar, F. . LyEmoBERT: Classification of lyrics' emotion and recommendation using a pre-trained model. Procedia Computer Science, 218, 1196-1208.(2023).
3. Shahriar, Sakib, and Noora Al Roken. "How can generative adversarial networks impact computer generated art? Insights from poetry to melody conversion." International Journal of Information Management Data Insights 2, no. 1 (2022): 100066.
4. Gaonkar, P., & Fernandes, A. R. . Digitization of Konkani texts, and their transliteration: An initiative towards preservation of a language culture.(2019)
5. Lee, Cheng-Che, et al. "Crossing you in style: Cross-modal style transfer from music to visual arts." Proceedings of the 28th ACM international conference on multimedia. (2020).
6. Lee, C.C., Lin, W.Y., Shih, Y.T., Kuo, P.Y. and Su, L.,, October. Crossing you in style: Cross-modal style transfer from music to visual arts. In Proceedings of the 28th ACM international conference on multimedia (pp. 3219-3227),(2020).

7. Shahriar, S.. GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network. Displays, 73, 102237.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982]. (2022)
8. Zhu, Y., Wu, Y., Olszewski, K., Ren, J., Tulyakov, S., & Yan, Y. Discrete contrastive diffusion for cross-modal music and image generation. arXiv preprint arXiv:2206.07771. (2022)
9. Zhao, Shihao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K. Wong. "Bridging different language models and generative vision models for text-to-image generation." In European Conference on Computer Vision, pp. 70-86. Cham: Springer Nature Switzerland, 2024.
10. Ng, Brian Man-Kit, Samantha Rose Sudhoff, Haichang Li, Joshua Kamphuis, Tim Nadolsky, Yingjie Chen, Kristen Yeon-Ji Yun, and Yung-Hsiang Lu. "Visualize Music Using Generative Arts." In 2024 IEEE Conference on Artificial Intelligence (CAI), pp. 1516-1521. IEEE, (2024).
11. Elhoseiny, A.E.B.L.M. and Mazzone, M., CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms.
12. Dhore, M., Dixit, S., & Dhore, R.. Optimizing transliteration for Hindi/Marathi to English using only two weights. In Proceedings of the First International Workshop on Optimization Techniques for Human Language Technology (pp. 31-48), (2012, December)
13. Yang, Meng, Maria Teresa Llano, and Jon McCormack. "Exploring Real-Time Music-to-Image Systems for Creative Inspiration in Music Creation." arXiv preprint arXiv:2407.05584 (2024).
14. Sharma, Hardik, Shelly Gupta, Yukti Sharma, and Archana Purwar. "A new model for emotion prediction in music." In 2020 6th International Conference on Signal Processing and Communication (ICSC), pp. 156-161. IEEE, 2020.
15. Gómez, Lucía Martín, and María Navarro Cáceres. "Applying data mining for sentiment analysis in music." In Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection-15th International Conference, PAAMS 2017 15, pp. 198-205. Springer International Publishing, (2018).
16. Choi, Joong Ho, Geonyeong Choi, Ji Eun Han, Wonjin Yang, and Zhi-Qi Cheng. "Music2P: A Multi-Modal AI-Driven Tool for Simplifying Album Cover Design." In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 5200-5204. 2024.
17. Nakatsuka, T., Hamasaki, M., & Goto, M. . Content-based music-image retrieval using self-and cross-modal feature embedding memory. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2174-2184).(2023)
18. Nakatsuka, T., Hamasaki, M. and Goto, M., Content-based music-image retrieval using self-and cross-modal feature embedding memory. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2174-2184).(2023).
19. Ha, David, and Douglas Eck. "A neural representation of sketch drawings." arXiv preprint arXiv:1704.03477 (2017).
20. Li, W., & Li, J. . Research on music visualization based on graphic images and mathematical statistics. IEEE Access, 8, 100652-100660.(2020)
21. Tan, Xiaodong, Mathis Antony, and H. Kong. "Automated Music Generation for Visual Art through Emotion." ICCC. 2020.
22. Han, Chi, et al. "Learning shared semantic space for speech-to-text translation." arXiv preprint arXiv:2105.03095 (2021).
23. Grzywczak, Daniel, and Grzegorz Gwardys. "Audio features in music information retrieval." International conference on active media technology. Cham: Springer International Publishing, (2014).
24. Tabaza, Abdulrahman, Omar Quishawi, Abdelrahman Yaghi, and Omar Qawasmeh. "Binding Text, Images, Graphs, and Audio for Music Representation Learning." In Proceedings of the Cognitive Models and Artificial Intelligence Conference, pp. 139-146. 2024.
25. Sun, Wei, and Revathi Sundarasekar. "Research on pattern recognition of different music types in the context of AI with the help of multimedia information processing." ACM Transactions on Asian and Low-Resource Language Information Processing (2023).
26. Oppenlaender, J.,, November. The creativity of text-to-image generation. In Proceedings of the 25th international academic mindtrek conference (pp. 192-202),(2022)

27. de la Fuente, Carlos, Jose J. Valero-Mas, Francisco J. Castellanos, and Jorge Calvo-Zaragoza. "Multimodal image and audio music transcription." International Journal of Multimedia Information Retrieval 11, no. 1 (2022): 77-84.
28. Özgen, A. C., Aghdam, O. A., & Ekenel, H. K. . Text-to-Painting on a Large Variance Dataset with Sequential Generative Adversarial Networks. In 2020 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.(2020, October)
29. Williams A, Lattner S, Barthet M. Sound-and-Image-informed Music Artwork Generation Using Text-to-Image Models.
30. Kim, Doyeon, Donggyu Joo, and Junmo Kim. "Tivgan: Text to image to video generation with step-by-step evolutionary generator." IEEE Access 8 (2020): 153113-153122.
31. Verma, G., Dhekane, E.G. and Guha, T., 2019, May. Learning affective correspondence between music and image. In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3975-3979). IEEE.(2019)