Title

Image to Music: Cross-Modal Melody Generation Through Image Captioning

Authors

Alper Kaplan[1], Dionysis Goularas[2*]

Affiliations

[1]Cognitive Science Department, Graduate School of Social Sciences, Yeditepe University, Istanbul, 34755, Turkey

[2]Computer Engineering Department, Graduate School of Natural and Applied Sciences, Yeditepe University, Istanbul, 34755, Turkey

*To whom correspondence should be addressed; E-mail: goularas@cse.yeditepe.edu.tr

Abstract

Advances in machine learning in recent years have also been seen in computationally creative systems. Interest in machine-generated artifacts paved a way for creative models to evolve as such. But the earlier methods mostly explored a one domain approach and cross-modal learning has stayed relatively unexplored. Thus, the direct mapping between modalities for cross-modal creative models is not fully explored. This work proposes a novel methodology for generating symbolic music through images by directly mapping their features. A CNN encoder and deep-stacked LSTM decoder are the base models as the proposed method uses the image captioning approach to map the two domains' features. The generated music is evaluated quantitatively by using a custom genre classification model and BLEU scores calculations. The qualitative evaluation involves a melody listening test with human evaluators. The results show that the proposed method works well for music generation.

Keywords

Music Generation, Melody Generation, Cross-Domain Learning, Image Captioning, Machine Learning, Deep Learning

INTRODUCTION

Humans are naturally talented and complex creatures. Thanks to their well-developed prefrontal cortex, they have the ability of higher cognition, which includes having episodic memory, self-awareness, and a theory of mind. They can recognize people's faces and objects; can understand and speak natural languages subconsciously. However, these seemingly mundane tasks to humans can be challenging for machines. Nonetheless, artificial general intelligence (AGI) is the ultimate goal in machine learning (ML) research: the hypothetical ability of a machine to perform any task that a human can undertake. Besides, one of the most significant challenges on the path to AGI is achieving human-level creativity.

Mel Rhodes (1961) describes creativity as *the phenomenon in which a person communicates a new concept*. This *new concept* can be anything from tangible, an invention of a device, a painting, writing a poem, or writing this thesis; to intangible, solving a problem, expressing an idea, making connections between apparently unrelated subjects, or thinking about the future. This research focuses on composing music as *new concepts*, not by *a person* but by *a machine*.

The past decade showed a high rise in ML-based techniques, spanning from image classification and object detection to natural language processing (NLP) for sentiment analysis and topic clustering tasks. Computationally creative intelligent systems were also rising with deep art and music generation research. However, converting data from one domain to another is relatively unexplored for music generation. This research proposes a methodology for cross-modal (image-to-) music generation.

The introduction is based on the idea of human/machine creativity, the relationship between music and language, and cross-modal learning. Starting with the next chapter (2. Literature Review), the rest of this thesis focuses on the technical details of the related machine learning research, dataset building, and developed models to achieve the desired output, results, and discussions.

Computationally Creative Systems

The capacity to create or develop novel work, theories, procedures, or ideas is defined as creativity. In this regard, computationally creative systems (CCSs) aim to achieve human-level creativity in their output artifacts. Although machines have not yet reached the competence to pass humans on creativity, researchers are trying to close this gap by working on CCSs. One research focuses on machine-generated art and emphasizes a computationally creative system (Heath & Ventura, 2016). By increasing the perceptual ability of the system, researchers showed that this resulted in better feature extraction from any given dataset (paintings). This approach adds more variety to generated artifacts, thus making them more unique and creative. Another research about art generation focuses on the generation and goes a bit further for being creative (Elgammal et al., 2017). Researchers proposed an intelligent system that can learn a painting style. Based on its experience, the system can generate novel artifacts using a modified Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) called Creative Adversarial Network (CAN). They also made a Turing test where people failed to distinguish between computer-generated and human-generated art.

On the other hand, music is mainly about statistics, making it a suitable study area for machine learning applications. One recent research proposed a genetic programming system using statistical and structural descriptors in a genetic system for melody CCSs 4 composition (León et al., 2016). Since the first machine-generated song, Lejaren and Leonard's Illiac Suite (1959), there have been many advancements in the technology used for machine learning, particularly neural networks, and their methodology for CCSs.

The deep learning technique for music generation gradually matured over the last twelve years. A project named DeepBach (Hadjeres et al., 2016) intended to model polyphonic music (hymn-like pieces). They used pseudo-Gibbs sampling coupled with an adapted representation of musical data which resulted in highly convincing chorales in the style of Bach. On the other hand, Zuckowski and Carr (2017) claimed that most style-specific generative music applications had explored artists commonly found in harmony textbooks, such as Johann Sebastian Bach, but very few researchers looked for song generation for modern genres such as black metal.

Building perceptually more aware systems in the sense of feature extraction and, in the area of generating/creating new artifacts, a combination of the presented features, such as machine learning algorithms and deep learning architectures of GAN/CAN, SampleRNN (Mehri et al., 2017), may yield with distinguished pieces (apart what is given as input from the dataset) in the sense of creativity.

Cross-Modal Learning: Image to Music

Cross-modal perception is a widely explored research topic in several disciplines, including psychology (Davenport et al., 1973; Storms, 1998, Vines et al., 2006; Vroomen & Gelder, 2000), neurology (Stein & Meredith, 1993), human-computer interaction (Mignot et al., 1993; Tanveer et al., 2015), and gained attention lately in computer vision (Owens et al., 2016; Owens & Efros, 2018) (see Figure 1), audition (Li et al., 2017) and multimedia analysis (Feng et al., 2014; Pereira et al., 2013). This work aims to develop an audio-visual generative machine learning architecture to generate meaningful (i.e., harmonic, consistent) melodies from input images. The challenge for this task presents itself as audio/visual feature extraction, cross-modality conversions, and conditional image/sound synthesis.

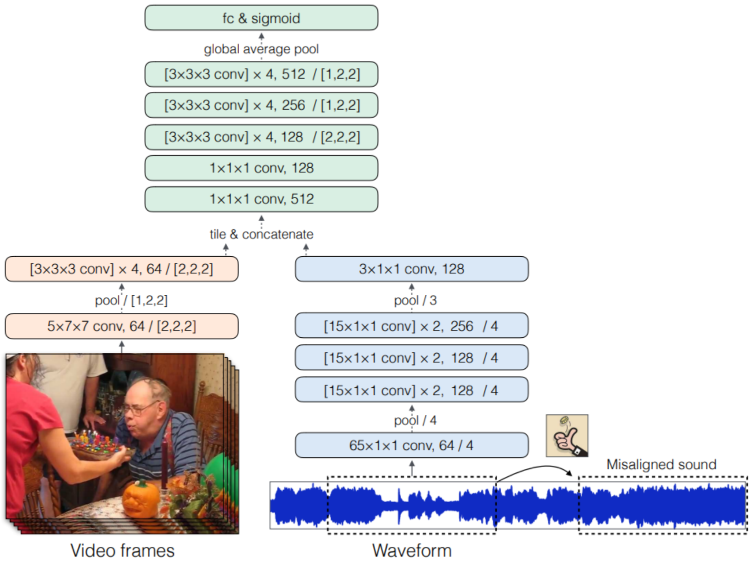


Figure 1: Proposed neural network model by Owens and Efros (2018). Using a fused multisensory representation, they state that a video signal's visual and audio components should be modeled jointly.

Although there are many works in cross-modal analysis, most research focuses on indexing and retrieval (Ngiam et al., 2011; Aytar et al., 2016; Arandjelovic & Zisserman, 2017) instead of generation. Moreover, when mapping one modality space to another, one of the most widely used generative architectures is GANs (Wan et al., 2019; Lyu et al., 2018). Recently, a cross-modal music generation model using GANs for image translation tasks was also proposed (Ruzafa, 2020).

Instead of using GANs, this work proposes a novel approach for cross-modality music generation based on image captioning with attention. Also known as automatic image annotation, a computer model automatically provides metadata to a digital image in captions or keywords. In computer vision, this is a task of scene understanding. Many successful applications emerged over the years by taking the image as a whole (Vinyals et al., 2015) and attending only to the relevant parts of an image (Xu et al., 2015). Image captioning tasks are usually based on the encoder-decoder architectures. The input of images is given to the architecture's encoder part, which is primarily a convolutional neural network (LeCun et al., 1998). Later, these features are combined with natural language data (captions) to be processed. A pattern is learned at the decoder part of the architecture, primarily an LSTM. Since image captioning is a developing research area, this brings another challenge to the issue of cross-modal symbolic music generation at hand.

METHODOLOGY

Dataset

Image Dataset

The image dataset consists of paintings from WikiArt (Saleh & Elgammal, 2015), a user-editable online visual art encyclopedia. It has over 120000 paintings where each artwork is annotated by its style, name, artist, date, and a URL containing the image file's location. These metadata are stored in a comma-separated value (CSV) file (kaggle & antoinegruson, 2022).

The images are crawled using their corresponding URLs by using the metadata information. Although WikiArt has images from 217 unique styles/eras, this work is only interested in a smaller subset of them. These are *Baroque*, *Classicism*, *Romanticism*, *Art-Nouveau (Modern)*, *Divisionism*, *Impressionism*, *Post-Impressionism*, and *Symbolism*. Apart from the first three styles, the rest of them are gathered under a custom-umbrella style called *Modern*.

Music Dataset

The music files come from the MIDI and Audio Edited for Synchronous TRacks and Organization (MAESTRO) (version 3.0.0) dataset (Hawthorn et al., 2018), which contains around 200 hours of virtuoso piano performances from Western classical music. The unique number of composers is 60, and it has 854 piano compositions. Each recording in the dataset has seven fields which are *canonical_composer*, *canonical_title*, *split*, *year*, *midi_filename*, *audio_filename*, and *duration*.

Table 1

The fields of the WikiArt dataset.

Field	**Description**
Style	Style of the artwork.
Artwork	Name of the artwork.
Artist	Name of the artist.
Date	Creation date of the artwork.
Link	Download link to the artwork as a JPG/JPEG or PNG image.

Note. From "the WikiArt dataset" by WikiArt.org

Table 2

The statistics of the MAESTRO v3.0.0 dataset.

Split	**Performances**	**Duration (hours)**	**Size (GB)**	**Notes (millions)**
Train	962	159.2	96.3	5.66
Validation	137	19.4	11.8	0.64
Test	177	20.0	12.1	0.74
Total	1276	198.7	120.2	7.04

Note. From "the MAESTRO dataset" by Magenta, TensorFlow, Google.

Preprocessing

Image Dataset Preprocessing

Image-text pairs associate image features with their corresponding captions for the image captioning task. For image-to-symbolic music generation, this work follows a similar fashion with an important distinction. The pairing process of images (paintings) and texts (symbolic music) is being done arbitrarily. It is certain that arbitrarily pairing images and symbolic music already brings much noise to the general distribution of the dataset. Therefore, it is quite important to have samples from the same distribution. For this reason, an anomaly/outlier removal procedure was followed using image features and superpixel-based images. Note that the outlier detection topic is out of the scope of this work. Hence, the only focus is on explaining the outlier removal process.

An outlier is an observation in a set that deviates too much from the rest of the points in that set. For this work, the used outlier removal algorithm is the isolation forest (Liu et al., 2008). It is a unique anomaly detection method that relies on isolation (the distance between a data point and the rest of the data) rather than modeling normal points. The aim is to make all the images as much as from the same distribution for each *style*. The custom umbrella style, *Modern*, has images from the sub-styles of *Art-Nouveau (Modern)*, *Divisionism*, *Impressionism*, *Post-Impressionism*, and *Symbolism*. Hence, each style has gone through the same outlier removal process. Note that histograms[1] of HSV[2] images are calculated and normalized before applying the isolation forest.

The outlier removal algorithm is first applied to raw images to remove the 10% of data from each *Style*. Then, image features are extracted using a recently proposed image classification model called ViT-L/16 (Dosovitskiy et al., 2021). This is an application for a vision transformer (ViT) (Vaswani et al., 2017) model, which is also used as an encoder in the proposed model's training phase. The details of the encoder are shared in the following chapters.

After feature extraction, the isolation forest algorithm is again applied to remove the 20% data from each *Style*. Later, the remaining images are converted into superpixel-based versions by using SLIC segmentation. Superpixels are the outcome

[1] Histogram: a graphical representation that shows the approximate distribution of numerical data.

[2] HSV: An alternative way to represent RGB images.

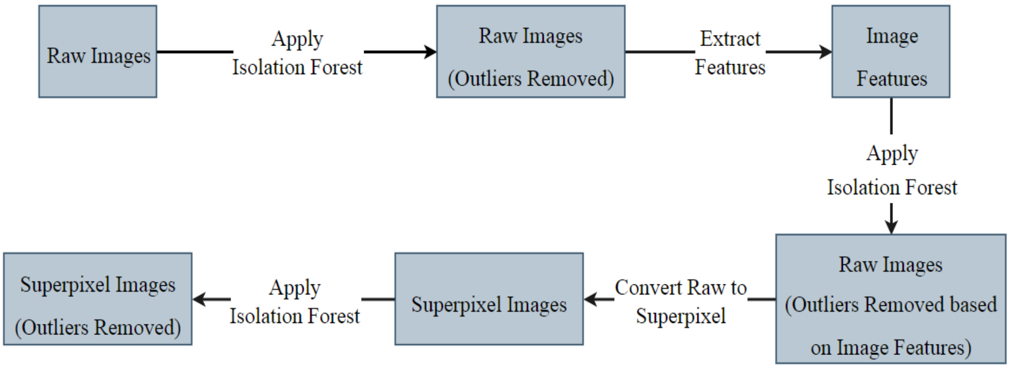

Raw Images
Apply Isolation Forest
Raw Images (Outliers Removed)
Extract Features
Image Features
Apply Isolation Forest
Raw Images (Outliers Removed based on Image Features)
Convert Raw to Superpixel
Superpixel Images
Apply Isolation Forest
Superpixel Images (Outliers Removed)

Figure 2: Outlier removal process from the image dataset.

of perceptual pixel grouping or the effect of picture over-segmentation. Superpixels store more information than pixels and better match picture borders than rectangular image patches. These converted images are also used in the training process. This has allowed having the gist (superpixel patterns) of each *Style* to effectively represent their respective color palette, edges, contours, and other related features. Recent studies showed that having superpixel-based image features results in better performance for multispectral image classification (Liu et al., 2017; Zhao et al., 2017). The examples of the superpixelated versions of the raw images can be seen.

Lastly, 60% and 10% of data were removed from *Modern* and from the rest of the data in each *Style*, respectively.

Arbitrarily pairing image and music files yield 2425 pairs for *Baroque*, 187 pairs for *Classic*, 2338 pairs for *Romanticism*, and 831 pairs for *Modern*. 95% and 5% of data split as the train and test sets, respectively. In total, there are 5451 training samples and 330 test samples.

Table 3

The statistics of the selected styles of the WikiArt dataset used in this work. From the number of raw files to the number of files after each preprocessing step applied.

Style	size			
	Raw	Raw (10% of outliers removed)	Features extracted (20% of outliers removed)	Superpixels (60% and 10% of outliers removed from Modernism and from the rest, respectively)
Baroque	3600	3238	2590	2425
Classicism	276	248	198	187
Romanticism	3600	3240	2592	2338
Modernism	14816	13330	10663	3991
Art-Nouveau (Modern)	3600	-	-	-
Divisionism	421	-	-	-
Impressionism	3600	-	-	-
Post-Impressionism	3595	-	-	-
Symbolism	3600	-	-	-

Table 4

Examples of images from their original versions from the WikiArt dataset (left) vs. their superpixelated versions (right).

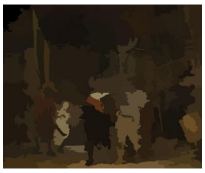
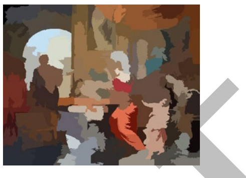
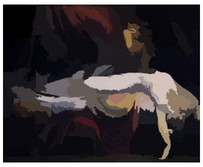
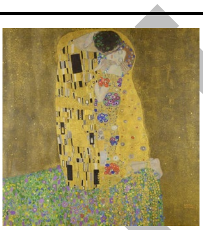
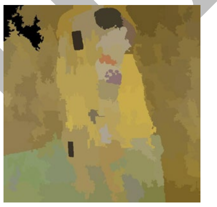
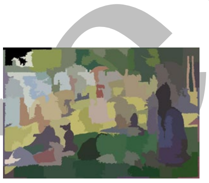
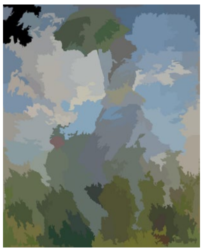
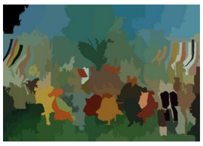
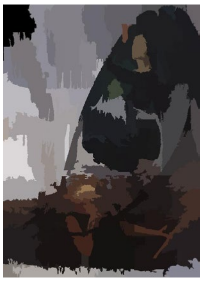
Baroque	Classicism
	
Romanticism	Art-Nouveau (Modern)
	 
Divisionism	Impressionism
	
Post-Impressionism	Symbolism
	

Table 5

The statistics of the used dataset for image-symbolic music pairs for each style.

Style	size	
	Training	Test
Baroque	2304	121
Classical	178	9
Romanticism	2221	117
Modern	748	83
Total	5451	330

Music Dataset Preprocessing

Since pairing paintings and symbolic music is based on art movements; the music has to come from the same art movement to make pairing semantically correct in an abstract manner. Consequently, the selected styles are *Baroque*, *Classical*, *Romanticism*, and *Modern*. The base comes from the Cross-Era (audiolabs-erlangen, 2022) dataset to distinguish each canonical composer with its respective style. In this work, their styles are laid out chronologically (Figure 3). There are composers in-between two styles, which are called *transitional composers*. They are not included in this work.

The MIDI files in the MAESTRO dataset are separated accordingly to each composer's style. After the separation process, each MIDI file is split into 5-second mini-recordings (melodies) to improve the quality of the proposed music generation method through image captioning. These melodies are preprocessed to increase the reliability of the image processing task by holding only the melodies with a certain number of notes.

Table 6

The metadata files in the Maestro dataset have seven fields for every MIDI/WAV pair.

Field	**Description**
canonical_composer	Composer of the piece.
canonical_title	Title of the piece.
split	Suggested train/validation/test split.
year	Year of performance.
midi_filename	MIDI filename.
audio_filename	WAV filename.
duration	Duration in seconds, based on the MIDI file.

Table 7

Statistics of the used dataset, before and after the split operation.

Style	size		
	Original (before split)	After split	After preprocessing
Baroque	165	11719	
Classical	117	9434	
Romanticism	894	110890	
Modern	7	748	

The symbolic music information is extracted from all the remaining music files by using the pretty-midi Python package. Note events can be obtained from a MIDI file by using this package. There are four events for each note: velocity, *pitch*, *note start time*, and *note end time*; e.g., the note events of a song from the MAESTRO dataset. To decrease the vocabulary size of the decoder model during training, one should keep in mind that the velocity events from notes are discarded.

The MIDI encodings are treated as *captions* for image captioning. The encodings include information such as *tempo* and *note events*. Also, each note event is treated as a *word* as if they were in a natural language. Hence, they are concatenated with an underscore (_) character. These encodings build the vocabulary. The embedding part of the decoder and generating new captions use these words in the vocabulary as reference.

Table 8

The events of the first ten notes from "Prelude and Fugue in E-flat Major" by Johann Sebastian Bach.

Velocity	Pitch	Start	End
42	G4	1.5	1.8
53	G#4	1.8	2.1
60	A#4	2.1	2.3
64	G#4	2.3	2.6
66	G4	2.5	2.8
64	F4	2.8	3.1
40	G3	3.8	4.1
50	G#3	4.1	4.3
52	A#3	4.3	4.6
52	G#3	4.6	4.9

Note. Each numeric value is rounded to 1 decimals.

Table 9

The tempo and the first 10 note events from the MIDI file of the third split of the composition, "Prelude and Fugue in E-flat Major" by Johann Sebastian Bach.

MIDI Event	**Value**
Tempo	230
Notes	
#1	D5_0.01_0.29
#2	D#5_0.26_0.5
#3	F5_0.0_0.74
#4	F5_0.49_0.74
#5	D#5_0.74_0.99
#6	D5_0.99_1.23
#7	C5_1.23_1.48
#8	D4_0.0_2.4
#9	D4_2.14_2.4
#10	D#4_2.4_2.66

RESULTS

Quantitative Results

As stated earlier, quantitative results are based on the custom genre classification model and BLEU score calculations. The genre classification model is applied as four classes (styles) together, namely baroque, classical, romanticism, and modern. They are also examined as paired classes which are baroque-classical, baroque-romanticism, baroque-modern, classical-romanticism, classical-modern, and romanticism-modern. Accuracy, precision, recall, and f1 scores are calculated for these predictions.

Table 10

Genre classification model results.

	accuracy (%)	precision (%)	recall (%)	f1 (%)
All classes				
Train (10 cross-validated)	81	83	83	83
Test (Human-generated)	83	83	83	83
Test (Machine-Generated)	43	43	43	43
Baroque-Classical				
Train (10 cross-validated)	96	97	97	97
Test (Human-generated)	97	97	97	97
Test (Machine-Generated)	89	89	89	89
Baroque-Romanticism				
Train (10 cross-validated)	90	93	93	93
Test (Human-generated)	92	92	92	92
Test (Machine-Generated)	52	52	52	52

Table 10

cont.

Baroque-Modern				
Train (10 cross-validated)	95	96	96	96
Test (Human-generated)	96	96	96	96
Test (Machine-Generated)	47	47	47	47
Classical-Romanticism				
Train (10 cross-validated)	95	95	95	95
Test (Human-generated)	95	95	95	95
Test (Machine-Generated)	93	93	93	93
Classical-Modern				
Train (10 cross-validated)	95	96	96	96
Test (Human-generated)	96	96	96	96
Test (Machine-Generated)	67	67	67	67
Romanticism-Modern				
Train (10 cross-validated)	88	88	88	88
Test (Human-generated)	88	88	88	88
Test (Machine-Generated)	75	75	75	75

The other quantitative results are based on BLEU scores. For each style, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores are calculated. The calculations are separated into two. First, the scores for each generated melody's whole sequence are used. The next one is the calculation based on only the notes of each generated melody.

Table 11

Generated melodies BLEU scores based on the training data.

	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**
Whole sequence				
Baroque	3.71	**23.0**	7.80	3.71
Classical	0	**3.83**	0	0
Romanticism	2.04	**15.04**	4.51	2.04
Modern	0.76	**3.92**	0.95	0.76
Only notes				
Baroque	94.66	**99.95**	99.79	94.66
Classical	57.28	**95.15**	86.96	57.28
Romanticism	95.23	**100**	99.93	95.23
Modern	77.35	**100**	98.16	77.35

Table 12

The results of the melody listening test with human evaluators.

	Turing test pass (%)	**Correct genre prediction (%)**
Baroque	**56**	59
Classical	37	60
Romanticism	52	**68**
Modern	46	70

Qualitative Results

The qualitative results are obtained by conducting a melody listening test with 10 human participants.

DISCUSSION

This work aimed to pave a map between two different modalities: image and music. Using the artifact pairs from the same art movement styles shows that creating a cross-modal generative model is possible. The adopted approach of image captioning also shows that images of paintings can be used to produce symbolic music sequences.

In the case of image captioning-based music generation, it is certain that having rich image features will lead to better results, meaning harmonious, coherent melodies. AI is a fast-growing field as a new SOTA image classification model is proposed almost every month. Changing the encoder with a SOTA model is definitely one of the options, e.g., the recently published model, CoCa (Yu et al., 2022). Also, changing the decoder's LSTM network with more up-to-date methods, such as with language transformers (Wolf et al., 2020), might also yield good results.

Alternative image segmentation to SLIC can also be adopted for the image segmentation part. Changing the hyperparameter of the *number of components* in the SLIC algorithm may also yield different results as they may capture the patterns of a painting, hence, the style.
Using different MIDI encoding techniques can be used for the decoder part as well, such as REMI (Huang et al., 2020), compound word transformer (Hsiao et al., 2021), structured (Hadjeres & Crestel, 2021), Octuple (Zeng et al., 2021), and MuMIDI (Ren et al., 2020). Additionally, these encodings, represented as text, can be converted to vectors, which might be used in the decoder part. Famous in NLP, word2vec is a text vectorization algorithm that is used to vectorize symbolic music (Chuan et al., 2020). Directly encoding the MIDIs is also an alternative. MIDI2vec (Lisena et al., 2020) is an example of this and can be used at the decoder part.

Since the first emergence of GAN, it has been excessively used for generative art tasks. For the image to music generation, image-to-image translation is an interesting approach proposed by Ruzafa (2020) using cGAN. It is one of the alternatives that can be examined in depth.

REFERENCES

Elgammal, A., Liu, B., Elhoseiny, M., Mazzone, M. (2017). Can: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. arXiv preprint arXiv:1706.07068.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

Ponce de León, P. J., Inesta, J. M., Calvo-Zaragoza, J., Rizo, D. (2016).Data-based melody generation through multi-objective evolutionary computation. Journal of Mathematics and Music, 10(2), 173-192.

Hadjeres, G., Pachet, F., Nielsen, F. (2017, July). Deepbach: a steerable model for bach chorales generation. In International Conference on Machine Learning (pp. 1362-1371). PMLR.

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., ... Bengio, Y. (2016). SampleRNN: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837.

Davenport, R. K., Rogers, C. M., Russell, I. S. (1973). Cross modal perception in apes. Neuropsychologia, 11(1), 21-28.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. Cognition, 101(1), 80-113.

Mignot, C., Valot, C., Carbonell, N. (1993, April). An experimental study of future “natural” multimodal human-computer interaction. In INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems (pp. 67-68).

Tanveer, M. I., Liu, J., Hoque, M. E. (2015, October). Unsupervised extraction of human-interpretable nonverbal behavioral cues in a public speaking scenario. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 863-866).

Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E. H., Freeman, W. T. (2016). Visually indicated sounds. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2405-2413).

Owens, A., Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 631-648).

Li, B., Dinesh, K., Duan, Z., Sharma, G. (2017, March). See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2906-2910). IEEE.

Feng, F., Wang, X., Li, R. (2014, November). Cross-modal retrieval with correspondence autoencoder. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 7-16).

Pereira, J. C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G. R., Levy, R., Vasconcelos, N. (2013). On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE transactions on pattern analysis and machine intelligence, 36(3), 521-535.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y. (2011, January). Multimodal deep learning. In ICML.

Aytar, Y., Vondrick, C., Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. Advances in neural information processing systems, 29.

Wan, C. H., Chuang, S. P., Lee, H. Y. (2019, May). Towards audio to scene image synthesis using generative adversarial network. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 496-500). IEEE.
Lyu, J., Shinozaki, T., Amano, K. (2018). Generating Images from Sounds Using Multimodal Features and GANs.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. Z. A., Dieleman, S., ... Eck, D. (2018). Enabling factorized piano music modeling and generation with the MAESTRO dataset. arXiv preprint arXiv:1810.12247.

Liu, F. T., Ting, K. M., Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Zhao, W., Jiao, L., Ma, W., Zhao, J., Zhao, J., Liu, H., ... Yang, S. (2017). Superpixel-based multiple local CNN for panchromatic and multispectral image classification. IEEE Transactions on Geoscience and Remote Sensing, 55(7), 4141-4156.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y. (2022). CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv preprint arXiv:2205.01917.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations (pp. 38-45).

Huang, Y. S., Yang, Y. H. (2020, October). Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1180-1188).

Hsiao, W. Y., Liu, J. Y., Yeh, Y. C., Yang, Y. H. (2021). Compound Word Transformer: Learning to compose full-song music over dynamic directed hypergraphs. arXiv preprint arXiv:2101.02402.

Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T., Liu, T. Y. (2021). Musicbert: Symbolic music understanding with large-scale pre-training. arXiv preprint arXiv:2106.05630.

Ren, Y., He, J., Tan, X., Qin, T., Zhao, Z., Liu, T. Y. (2020, October). Popmag: Pop music accompaniment generation. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1198-1206).

Chuan, C. H., Agres, K., Herremans, D. (2020). From context to concept: exploring semantic relationships in music with word2vec. Neural Computing and Applications, 32(4), 1023-1036.

Lisena, P., Meroño-Peñuela, A., Troncy, R. (2020). MIDI2vec: Learning MIDI embeddings for reliable prediction of symbolic music metadata. Semantic Web, (Preprint), 1-21.