From Mondrian to Modular Synth: Rendering NIME using Generative Adversarial Networks

Akito van Troyer
Berklee College of Music
MIT Media Lab
Boston/Cambridge, MA, USA
avantroyer@berklee.edu

Rébecca Kleinberger
MIT Media Lab
75 Amherst St
Cambridge, MA, USA
rebklein@media.mit.edu

ABSTRACT

This paper explores the potential of image-to-image translation techniques in aiding the design of new hardware-based musical interfaces such as MIDI keyboard, grid-based controller, drum machine, and analog modular synthesizers. We collected an extensive image database of such interfaces and implemented image-to-image translation techniques using variants of Generative Adversarial Networks. The created models learn the mapping between input and output images using a training set of either paired or unpaired images. We qualitatively assess the visual outcomes based on three image-to-image translation models: reconstructing interfaces from edge maps, and collection style transfers based on two image sets: visuals of mosaic tile patterns and geometric abstract two-dimensional arts. This paper aims to demonstrate that synthesizing interface layouts based on image-to-image translation techniques can yield insights for researchers, musicians, music technology industrial designers, and the broader NIME community.

Author Keywords

Image translation, generative adversarial network, musical interfaces

CCS Concepts

•**Human-centered computing** → **Interface design prototyping;** •**Theory of computation** → *Adversarial learning;* •**Computing methodologies** → *Graphics systems and interfaces;*

1. INTRODUCTION

The ability to create New Interfaces for Musical Expression (NIME) has so far remained in the hands of humans and possibly of some animals. Though the fabrication, programming and musical potential of those interfaces are increasingly assisted by the use of computer systems such as computer music, computer-aided design (CAD) and computer-controlled fabrication machinery, the process of conceiving and envisioning the design of the interfaces is still a human task. In this paper, we propose to teach computers to automatically create new musical interface layouts using image-to-image translation techniques based on generative adversarial networks (GANs) [12]. Furthermore, the paper also examines resulting experimental layouts and suggest their potentials in aiding NIME builders' design process.

Figure 1: left: Abstract Composition by Erich Buchholz used as input image. right: resulting musical interface design generated using Model 3.

Most existing artificial intelligence (AI) based implementations in the realm of music aim at creating novel listening experiences. Indeed, in recent years, AI algorithms have gained considerable attention from music community in domains such as musical content generation and music information retrieval [1, 8, 13, 7]. In particular with content generation, most existing AI based implementations for music are focused on the manipulation of symbolic data (e.g. MIDI) or sub-symbolic data (e.g. audio signal). Using AI algorithms, musicians and technologists can now translate music across genres, styles, and musical instruments [21]. Other AI implementations merge the unique timbres of different instruments into new unheard sounds [10]. On a bigger scale, researchers have for a few decades worked on using machine intelligence to automatically generate entirely new musical pieces in the style of a specific composer starting with Cope's work in the 90's [5, 6] or for live improvisation based on artificial neural network [15]. In addition from creating new pieces, computers can also learn to improvise with performers in real-time [30]. In the domain of performance systems, projects like Wekinator [11] also use machine learning to learn to recognize user input gestures for music control.

Meanwhile, visual-based generative systems have improved grandly and have become mainstream. Introduced in 2014, Generative Adversarial Networks (GANs) [12] have opened the door to an entire wave of work in the visual domain. GANs, first introduced by Goodfellow et al., typically composed of two deep networks, a generative and a discriminative model, which compete with each other based on a game theory [12]. GANs are now widely used in image synthesis and editing applications because of their ability to produce realistic images. For example, GANs have been unsed for high-quality image generation [2], image blending [29], and

NIME'19, June 3-6, 2019, Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

image inpainting [24]. In particular, and of utmost interest for our work, is image-to-image translation, a constrained image synthesis technique using GANs. This technique synthesizes new images based on inputs such as images, texts, and sketches [16]. Some early high visibility work using GANs introduced generative artworks by transferring the style of a famous artist such as Monet and Van Gogh [31]. A new training model for GANs that separates high-level attributes and stochastic variation can also synthesize convincing human face images [17]. The technique has also been applied to transfer movements from one person to another in generated videos [3]. The power of image-to-image translation lies in the mapping functions that require less parameter tweaking, making the technique more popular.

Visual aesthetic and organizational layout are however also critical in the process of building musical interfaces and controllers. Musical and MIDI interfaces are pervasive and recognizable by their unique aesthetic often composed of knobs, buttons, sliders, and keyboards. Fels and Lyons detail the six steps to build a NIME as 1. Choose control space, 2. Choose sonic space, 3. Design mapping between control and sound output, 4. Assemble with hardware and software, 5. Compose and practice, 6. Repeat and Refine [19]. Moreover, the step 4 on assembling the hardware has critical consequences on the final musical result. As Perry Cook states: "the music we create and enable with our new instruments can be even more greatly influenced by our initial design decisions and techniques" [4]. There are thousands of existing audio interface to make music, maybe tens of thousands if including all custom designs and non-commercial instruments built by the NIME community. When technologists and researchers create a NIME, colors, texture, material, hardware, and layout are an important part of the design process. Those design choices are guided by ergonomic, aesthetic and musical rules but also by tradition and the creativity of the designers. Moreover, the visual design of those interfaces not only matters for the performer but also for the audience and the resulting holistic experience of performances [23].

In this project, we explore how AI could help researchers and musicians in creating NIME specifically in the process of laying out control components and choosing panel organization based not only on prior art but also on more abstract and elevated sense of aesthetic. To this end, we trained three models to be able to render an image's semantic content using different artistic and geometric styles. The results obtained are not meant to be taken literally to produce new commercial MIDI interfaces, but more as a proof of concept from which we could derive insights regarding human perception, forms of aesthetic, ergonomy and traditional designs. We were also curious to see if machines generate layouts that could create new objects that could not have been imagined by humans.

2. METHODOLOGY

Our goal is to produce new images of musical interfaces given existing images of musical interfaces and another set of images used for translation. We trained three different models. Model 1 based on the pix2pix approach and Models 2 and 3 based on the cycleGAN approach [16, 31]. pix2pix employs image-to-image translation based on conditional GAN (cGAN), a GAN model based on feeding auxiliary data to both discriminator and generator [20] to learn a mapping from an input image to an output image. The generator is trained to apply transformations to the input images to produce synthetic image outputs that may not be distinguishable from the original image inputs by the discriminator. The discriminator is trained to compare the input images from the generator to an unknown image and tries to detect the synthetic images. In this training process, the generator learns to fool the discriminator. In Model 1, we use pix2pix to explore the potential of translating outlines and hand-sketched musical interfaces into realistic image renderings of finished products.

Contrasting to pix2pix, cycleGAN (cycle-consistent adversarial networks) applies image-to-image translation without requiring paired images for training. cycleGAN uses two generators and two discriminators in the training phase. One generator is responsible for translating images from one domain to another while another does the inverse translation. Each generator has a corresponding discriminator that identifies synthetic images from the original in a similar manner to pix2pix. The advantage of using cycleGAN lies in its effectiveness in style transfer, including painting to photo transfer, collection style transfer, and season transfer. We used cycleGAN to generate two models: one that transfers mosaic tile patterns into musical interfaces (Model 2) and another generating interfaces from abstract geometric paintings (Model 3).

2.1 Databases

The models created in this paper were based on one database of images of musical interfaces (database A) and three different transfer databases (databases B1, B2, and B3). To render NIME using GANs we first created the interface database (Database A) to teach our system what a musical interface looks like. Database A contains front and top facing views of a variety of commercial MIDI keyboards, synthesizer modules, audio mixing console, samplers, drum machines, sequencers, and old style tape recorders. The 1120 images were gathered from sites such as Google images, Sweetwater, Yamaha, Roland, and other specialized music gear websites. The database was then cleaned for errors and duplicates, and the image size was normalized to 256x256 pixels.

We chose to use commercial musical interfaces as our database lies in the accessibility, abundance, consistency, and quality of the images. Photos of commercial musical interfaces are easily accessible through the Internet. their layouts have common components such as knobs, sliders, keyboards, pads, and switches. These interface components also have a long history in being used in the creation of electronic instruments. For instance, one of the first electronic instrument, telharmonium, was already equipped with a keyboard [28]. Because of this shared history, such interfaces often have consistent features. Concerning image quality, the photos of commercial musical interfaces are often taken in a controlled environment with a clean white background, ideal for model creation.

Model 1 with pix2pix is built to automatically generate convincing interfaces from simple outlines both to explore recurring qualities present in our interface database through generation of textures and color palettes, but also to generate good outcomes from hand-drawn sketches. We created a unique database of paired images A and B1 where B1 is obtained through the contour extraction of the original target image from database A. The images of outlines were then concatenated to the original images side by side to form a pair.

For Model 2, we explored the cycleGAN approach applied to database A and database B2 containing 613 photographs of mosaic tile patterns across different cultures such as French, Japanese, and Islamic. Our objective was to see how the style of mosaic pattern could be transferred onto new interface as such patterns are geometrically consistent while being aesthetically unique. Similarly to musical controllers, tiles have two-dimensional organizational qualities, a certain repetitivity and intricate layout that have been refined for centuries.

For Model 3, we again used style transfer but this time on database B3 containing 1037 geometrical abstract paintings from 51 different artists gathered from WikiArt [25] including Piet Mondrian, Alexander Rodchenko, and Camille Graeser. The pieces date from the early 19th century until today and the style range from Constructivism, Minimalism, Op Art, Neoplasticism, to Concretism. The choice of this database was motivated both by rational system-based arguments and by more philosophical ones. In order to find human-logic and aesthetic in the results, we needed rectangular and geometrical objects containing interesting patterns, color palettes and aesthetic sense of order to guide the generation of our novel interfaces. Besides, similarly to how NIME are a modern take on one of the most ancient art form, geometric abstract painters offer a relatively recent turn in the world of painting. Indeed, abstract geometric artists are masters in extracting the essence of beauty in a two-dimensional canvas and from minimal, and seemingly simple forms can lead the audience eyes into a fundamental experience. Taking the Bauhaus movement, for example, they have a vast knowledge in harmony and aesthetic that can sometimes lack in the design of musical interfaces [9].

3. RESULTS

3.1 Model 1: outlines

We tested Model 1 (B1 paired database of outlines and pix2pix approach) on 100 new images of outlines generated form new original target interfaces and on five hand-drawn sketches of imaginary musical interfaces.

Figure 2: Results from Model 1 using original target interfaces (left column: input images of interface outlines, middle column: result, right column: original target used to obtain the outline)

From the outline test, we compared each result to the original target image (see Figure 2). The generated images respected the outline and filled the image with various colors for the interface components. The generated and target images were compared using visual cues such as body, button, and keyboard colors. All the resulting images containing keyboards were correctly identified as such and populated 100% of the cases with white for the white keys and black for the black keys. Even in the case where the original keyboard colors were inverted (see Figure 4 row 5). 90% of the original interfaces had either black, grey or white body color and 80% of all the resulting images had correctly identified the interface body color. This could be partially explained by the outline extraction algorithm used, but not only as one resulting image correctly populated the interface body in brown similarly as the original target (see Figure 2 row 6). Most of the times, when a different body color was generated, the original target interface had nontraditional body color such as pink, yellow, red or blue. As for other interface components (buttons, knobs, sliders, etc.), the model filled them with generic colors not always related to the original target images. Regarding shape semantics, the model systematically interpreted circular outlines as patch plug input or knobs, and rectangular/square outlines as buttons that were more likely to be populated with bright color.

We also tested the model on five hand-drawn sketches of imaginary musical interfaces (see Figure 3). The resulting images were colored versions of the sketch populated with different colors for different elements. Even though the input images had no perfect straight lines, the model correctly interpreted most keyboard features to the point of correctly coloring even very curled keyboards. The resulting interface bodies were either grey or black, and circular outlines were either white or black.

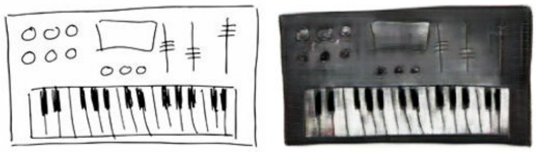

Figure 3: Results from Model 1 using hand drawn sketches (left : hand drawn input, right: result)

3.2 Model 2: Mosaic

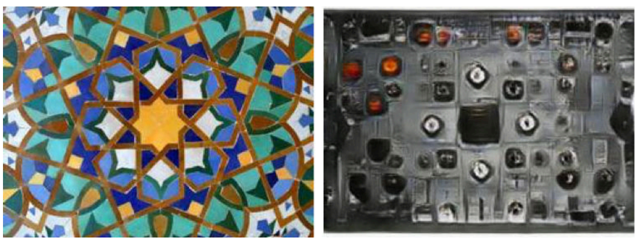

Figure 4: Results from Model 2 "B to A" (left: input image of tiles pattern. right: resulting musical interface layout design)

We tested Model 2 (B2 database of mosaic tiles and cycleGAN approach) both ways on a series of 24 new images of mosaic (B to A) and 24 new images of musical interfaces (A to B). The images generated from the "A to B" process did not present any aesthetic interests nor insights regarding the creation of new interfaces. Among the generated "B to A" images, about half could be easily recognized as control interfaces (see Figure 4 and 5). Five resulting images presented keyboard-like features, and 14 presented button-like features. Most square input images had been transformed into rectangular outputs which seems to indicate that the model tends to recreate a specific rectangular width to length ratio. Most resulting images (83%) had a white, black or grey dominant background and only a few color spots on the buttons zones.

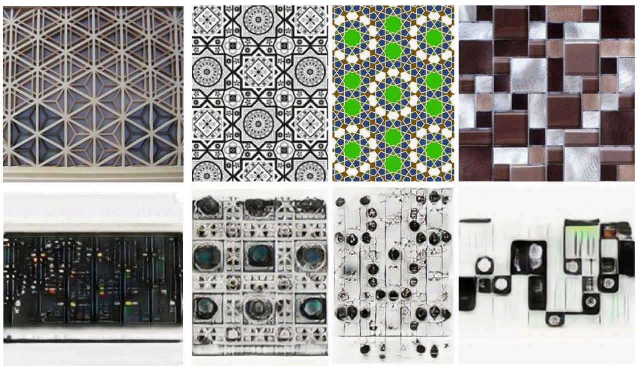

Figure 5: additional images obtained using Model 2 (top raw: input images, bottom row: results)

3.3 Model 3: Art

We tested Model 3 both ways (A to B and B to A) on a series of 173 new images of Artwork (B to A) corresponding to about four pieces for every 51 Artists. We also tested the model on 36 new images of interfaces (A to B) not contained in Database A.

Figure 6: Results from Model 3 on four paintings from Lajos Kassak (input image on the top row and result on the bottom row) (from left to right: Untitled; Constructivist Composition; Composition; Architectural Structures; Architectural Structures

Nearly all 173 resulting images from the "A to B" test present clear musical interface-like features (see Figure 8). Each is unique and present similarities with the original art piece used as its input. Only eight images (4%) contained keyboard-like features. Whereas most of the input images contain very bright colors, virtually all the resulting interfaces have a body color on a greyscale. All knobs, buttons and patch plug inputs are clearly defined and although their hue is generally in the blue or red domains, their saturation and brightness are uniquely coordinated with the color palette of the interface as a whole.

When looking at results from individual artists, we can observe even more consistency in the features generated by the model. The mapping of the colors seems to translate across paintings as well as a general balance of different interface components as seen in the series from Lajos Kassak (see Figure 6). When looking at the four interfaces generated from Mondrian paintings (see Figure 7), we can observe

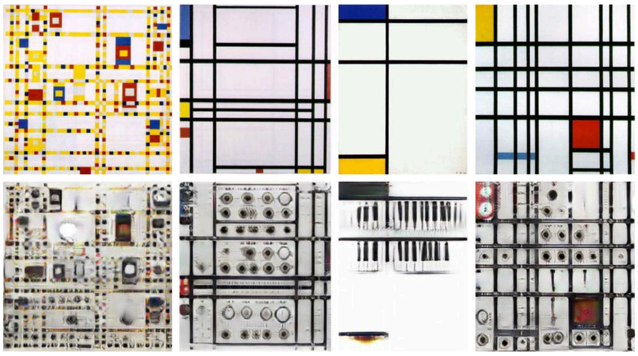

Figure 7: Results from Model 3 on four paintings from Piet Mondrian (input image on the top row and result on the bottom row) (from left to right: Broadway Boogie Woogie; Composition No.10; Composition III with Blue, Yellow and White; Composition with Red, Yellow and Blue

a direct relationship that the generated interface has to the original painting in terms of forms and shapes. Some of the colors are also respected. The empty spaces are generally recognized as panels and filled with modular synthesiser-like components.

3.4 Interface to Art

One final step performed with Model 3 was to generate 36 new artworks based on musical interfaces (A to B). Some of them seem closely related to the input interface and others have a less direct link. Each resulting image has their own unique aesthetic, shape scementic and color palette. Such result could be useful in audio visual work: artists working on a performance can base their visuals on their actual working interfaces so that audience members could better understand the musical piece.

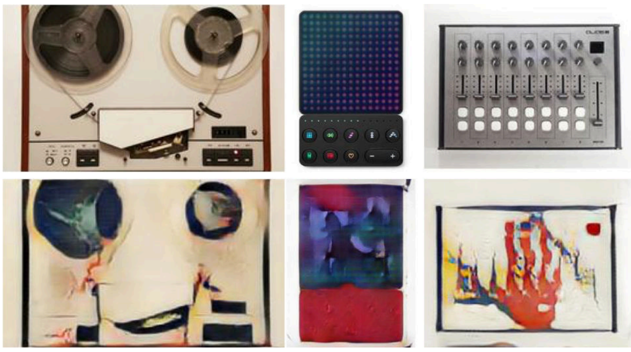

Figure 8: Results from Model 3 (top row: input interface image; bottom row: output artwork)

4. DISCUSSION

4.1 Insights

In the previous work using GANs, researchers have evaluated their work by assessing how visually convincing the result of translation is using platforms such as Amazon Mechanical Turk (AMT) [26]. In our case, we are not trying to produce photorealistic images nor to actually automatically generate new interfaces by putting the human out of the loop. Instead, we are proposing to gain insight into the current NIME design method by turning the entire process inside out and injecting new perspectives through the practice demonstrated in this paper. Our project first acknowl-

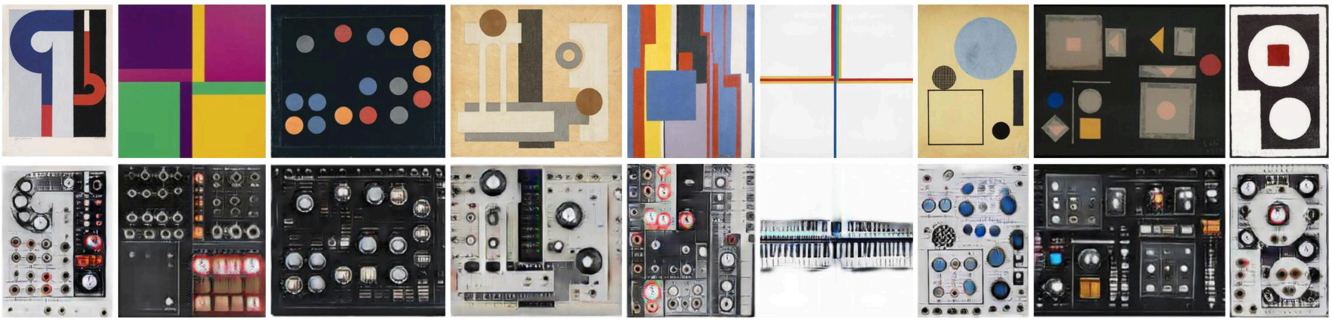

Figure 9: Results from Model 3. Top row: input images of Artworks from Yvan Serpa, Richard Paul Lohse, Lidy Prati, Otto Gustav Carlsund, Lothar Charoux, Max Bill, Henryk Berlewi, Lolo Soldevilla, Erich Buchholz). Bottom row: results

edges the existing beauty and inherent visual aesthetic of musical interfaces. Furthermore, by taking their aesthetics as a starting point to create new visuals to ultimately lead to new musical expressivity. Our contribution lies in the visual interest of our current results and the provocation of an ultimate form of musification. By letting an enigmatic visual become a guide for a possible mapping, itself inspired by seemingly random layout but actually led by another form of inspiration, in a way as visceral as music but in the visual domain.

We do also believe that our approach could lead to some more tangible future outcomes. In the commercial domain, for instance, this approach can open new avenues for artists to collaborate with music companies to create new products based on their visual artworks. Imagine architects, biologists, voice actors, and blacksmiths coming together to build the next generation of musical interface using GANs. This could become a way to expand the NIME community to include people from different backgrounds to collaborate in the music making process.

4.2 Future Directions

In the future, several variations could enrich our existing three models. To establish the full potential of a pix2pix approach, we plan to compare the results when using different edge detection algorithms to better simulate hand-sketched quality which might yield better results. In the introduction, we justified the use of abstract geometric two-dimensional art for the training of Model 3 but in the future work, it could also be interesting to train a model with other types of art forms from rupestre paintings to NaÃŕve Art or Cyber Art.

Another future step would be to extend our current interface database (database A) to include more original New Interfaces for Musical Expression created in the base few decades by the NIME community. Indeed, one could argue that less traditional interfaces such as the one presented at NIME could yield more interesting results. However, creating a database based on all previous NIME can be more challenging because their photos are taken in various environments with different qualities. Furthermore, images of NIME are not easily accessible and their visual appearance significantly varies and is less consistent. The first step for such an initiative might be to create an extensive inventory as well as guidelines for the community to document and photograph their work to be incorporated into a large open source catalog.

A plan to conduct a visual perception study using on AMT as a form of evaluation is already on the way. Not as a way to quantify how realistic the interface looks, but to further provoke the idea of turning the design process inside out and letting people imagine what such an instrument could sound like and be played.

Through AMT, having human observers evaluating the visual plausibility of interfaces rendered through GAN will further justify the contributions of our approach in the design process of NIME in the future. Based on the AMT visual rating from the human observers, we would also like to physically build one of the generated interfaces. This will enable us to evaluate the effectiveness of rendered musical interface images from an ergonomics perspective. Some Human-Computer Interaction (HCI) studies suggest that the visually appealing interfaces are more usable than those that are not [27, 22]. However, as HCI and Human Factors (HFs) have traditionally been concerned about usability more than the aesthetic [18], conducting a usability testing on a physical interface will further validate the use of GANs in the process of prototyping a musical interface hand-in-hand with humans.

Building the physical version of a generated image would raise a number of other questions. For example, how do we map the layout, color, and size of the interface components the resulting sound? Mapping is a subject dear to the NIME community and often interpreted as a gesture to sound mapping [14]. To our knowledge, this work sways the question towards thinking about mapping without the gesture: how does interface components themselves map to sound? For us, it seems that some of our current results already call for certain sonorities and interaction patterns, but those still need to be explored more in depth. As for this matter, we are also considering to run studies on AMT to assess how the âĂIJlook and feelâĂİ of an interface stimulates the imagination of observers in terms of music and sound. Our current visual results already shed light on some existing practices in musical interface design that one could question: what guide the color palette of commercial interfaces? What is the intrinsic visual language of modular synthesizer modules?

5. CONCLUSIONS

This paper explored the potentials of image-to-image translation techniques in aiding the design of new hardware-based musical interfaces. In this process, we collected a large set of images and trained three different generative models. Based on the generated images, we discussed to what extent GANs can give insights about the design process of new interfaces for musical expression to researchers, musicians, music technology industrial designers, and the broader NIME community.

6. REFERENCES

[1] J.-P. Briot, G. Hadjeres, and F. Pachet. Deep learning techniques for music generation-a survey. *arXiv preprint arXiv:1709.01620*, 2017.

[2] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[3] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018.

[4] P. Cook. Principles for designing computer music controllers. In *Proceedings of the 2001 conference on New interfaces for musical expression*, pages 1–4. National University of Singapore, 2001.

[5] D. Cope. Pattern matching as an engine for the computer simulation of musical style. In *ICMC*, 1990.

[6] D. Cope and M. J. Mayer. *Experiments in musical intelligence*, volume 12. AR editions Madison, 1996.

[7] L. Deng, D. Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[8] J. S. Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003.

[9] M. Droste. *Bauhaus, 1919-1933*. Taschen, 2002.

[10] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1068–1077. JMLR. org, 2017.

[11] R. Fiebrink and P. R. Cook. The wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, 2010.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[13] P. Hamel and D. Eck. Learning features from music audio with deep belief networks. In *ISMIR*, volume 10, pages 339–344. Utrecht, The Netherlands, 2010.

[14] A. Hunt, M. M. Wanderley, and M. Paradis. The importance of parameter mapping in electronic instrument design. *Journal of New Music Research*, 32(4):429–440, 2003.

[15] P. Hutchings and J. McCormack. Using autonomous agents to improvise music compositions in real-time. In *International Conference on Evolutionary and Biologically Inspired Music and Art*, pages 114–127. Springer, 2017.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[17] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[18] G. Lindgaard and T. A. Whitfield. Integrating aesthetics within an evolutionary and psychological framework. *Theoretical Issues in Ergonomics Science*, 5(1):73–90, 2004.

[19] M. Lyons and S. Fels. How to design and build new musical interfaces. In *SIGGRAPH Asia 2015 Courses*, page 9. ACM, 2015.

[20] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[21] N. Mor, L. Wolf, A. Polyak, and Y. Taigman. A universal music translation network. *arXiv preprint arXiv:1805.07848*, 2018.

[22] M. Moshagen, J. Musch, and A. S. Göritz. A blessing, not a curse: Experimental evidence for beneficial effects of visual aesthetics on performance. *Ergonomics*, 52(10):1311–1320, 2009.

[23] G. Paine. Towards unified design guidelines for new interfaces for musical expression. *Organised Sound*, 14(2):142–155, 2009.

[24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[25] F. Phillips and B. Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011.

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[27] A. Sonderegger and J. Sauer. The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied ergonomics*, 41(3):403–410, 2010.

[28] R. Weidenaar. *Magic music from the telharmonium*. Reynold Weidenaar, 1995.

[29] H. Wu, S. Zheng, J. Zhang, and K. Huang. Gp-gan: Towards realistic high-resolution image blending. *arXiv preprint arXiv:1703.07195*, 2017.

[30] M. Young. Nn music: improvising with a 'living' computer. In *International Symposium on Computer Music Modeling and Retrieval*, pages 337–350. Springer, 2007.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.