Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-10 (2022) 3292–3297

Using GANs to Generate Lyric Videos

Daniel Gareev * **Oliver Glassl** * **Sana Nouzri** *

* *University of Luxembourg, 2 Av. de l'Universite, 4365 Esch-sur-Alzette*

Abstract:
Artificial intelligence (AI) technologies have become increasingly common in creative practices in recent years. The rising number of research initiatives that emerge at the intersection of AI and art prompts researchers and artists to analyze the creative and explorative applications of AI in the context of art. First, this paper describes a specific AI art piece, an AI-generated music video for a song called Initiation, illustrating how AI can be used for creative purposes. This art piece was created in the context of the opening of Esch2022, the European Capital of Culture. Then, the paper provides an overview of the potential implications of AI technologies on the general understanding and creation of art.

Keywords: Artificial intelligence, art, expressive artificial intelligence, generative adversarial networks, image generation, video generation

1. INTRODUCTION

In 2022, Esch-sur-Alzette in Luxembourg is set to become the European Capital of Culture (Esch2022) (Wikipedia (2022a)). In this context, the University of Luxembourg hosts the AI & Art Pavilion, a multidisciplinary initiative trying to reflect the emerging future of artificial intelligence (AI) in the context of art. As a way to contribute to this mission, we explored various ways of using AI to create a music video for *Initiation* (Glassl and Savo (2019)), a song composed and produced by the *ThalamusProject* (Glassl and Savo (2022)). The *ThalamusProject's* name refers to the thalamus, a neural structure located in the human brain. It is frequently referred to as the *"gateway to consciousness"* or *"gateway to the mind"* (Ward (2013)), as it relays sensory information to the human consciousness (Torrico and Munakomi (2019)). The music of the *ThalamusProject* directly relates to this picture: acoustic textures are used as a canvas to depict the interplay of human subjectivity and objectivity. The human inner experience of external stimuli is hereby of special interest. The purpose of the experiment was to create a cinematic vision of the song from within the consciousness of an AI. The GAN model was trained on thousands of images collected online which depict the conceptual meanings of the song's lyrics. As it learned, the model developed neural models representing these concepts. The creators' final edit takes the spectators through the multidimensional latent space of the neural models, traveling across morphing landscapes and newly created artworks. This paper will discuss the implications of expressive AI practice by inductively learning from our AI-based artwork. We will first cover both the creative and technical aspects of the AI-based art piece. This will act as a concrete example to guide the rest of the discussion. Then, we will provide an outlook on the future development of AI technologies in the context of art.

1.1 Problem Description

Recent advances in AI have been driven by the development of novel algorithms and an increase in the availability of large amounts of data. This prompted the exploration and application of AI in various domains and raised several concerns on *"lack of interpretability, the limits of machine intelligence, potential risks, and social challenges"* (Cetinic and She (2022)). Across multiple discourses on AI, possibly the most equivocal one is the creation and understanding of art. The current technological advancements in AI have transformed the way we not only produce but consume art. Traditionally, artworks were crafted manually by humans. However, with the extensive development of machine learning (ML), it has now become possible to generate human textual, visual, or musical creations without human intervention. Among many forms of art, music and video have become the dominating content on social platforms. However, creating video clips to accompany music can be challenging for some creators due to limited budget, copyright constraints, search engine limitations, and required video-editing skills. One possible approach to offer artists more creative inspiration and the ability to explore AI in the context of art is to use AI to generate music videos. Most of the non-instrumental music contains lyrics. Therefore, one question naturally arises: What happens when a song's lyrics are used to create music videos? In this experiment, we explored the potential of AI to create visuals (e.g., images) by translating linguistic content and its semantics (e.g., song lyrics) into visual content. We thereby presumed that the result could reflect how AI, if a conscious entity, would understand and experience the ideas conveyed by the song's lyrics based

10.1016/j.ifacol.2022.10.126

on visual representations (e.g., images) that are sourced on the internet.

2. RELATED WORK

The creative applications of generative adversarial networks (GANs) have been explored widely in the last few years. For example, recent work on text-conditioned image generation (e.g., models that can generate images from text) has opened up many creative applications of GANs. DALL·E from OpenAI can create convincing images for a variety of sentences that *"explore the compositional structure of language*" (Ramesh et al. (2021)). The recently released version of DALL·E 2 produces even higher-quality samples and is computationally more efficient (Ramesh et al. (2022)). However, the model has not yet been released to the public. In the most similar work to ours, Frosst et al. (2019) used text conditional GAN (AttnGAN) and generated an image for every lyric line of a song. Then, they used each line in the song as the input for the model, thereby creating a series of images that represent the lyrics. To turn the series of images into a video, they interpolated between the images in time with the music. However, they used a pre-trained model trained on a publicly available dataset. This might be a limiting factor to some creators as they might not have enough artistic freedom to convey the visual narrative of the music. Even if the models are not pre-trained, it is still difficult to train the text-to-image generative models as the users need to collect both the images and text pairs as features. Additionally, the AttnGAN (Xu et al. (2018)) model is only able to produce images 256x256 pixels in resolution. In contrast, some of the recent models, such as StyleGAN2-ADA (Karras et al. (2020a)), are considered the state-of-the-art image generative models capable of producing images 1024x1024 pixels in resolution.

In this work, rather than using pre-trained text-to-image models, we propose an approach for leveraging existing state-of-the-art GANs to generate videos to accompany the music. Previous research in art and AI has also often focused on generating a single modality (e.g., visuals) without considering the context of other modalities (e.g., text). In this paper, we would like to explore the relationship between multiple modalities (e.g., text, visuals). Specifically, we would like to explore how a specific modality (e.g., lyrics of a song) can be used to create another modality (e.g., images). Naturally, the question arises: How will this newly created modality relate to the original modalities (e.g., lyrics and music)?

3. METHODS

The creation of the lyric video is a process involving multiple steps:

(1) We sourced the images from a search engine using the search queries that represent the semantic concepts of the song's lyrics.
(2) We clustered the images into n groups to select the photos that represent the semantic concepts.
(3) We trained a GAN-based model on the collected images and generated a number of interpolations between the pictures.

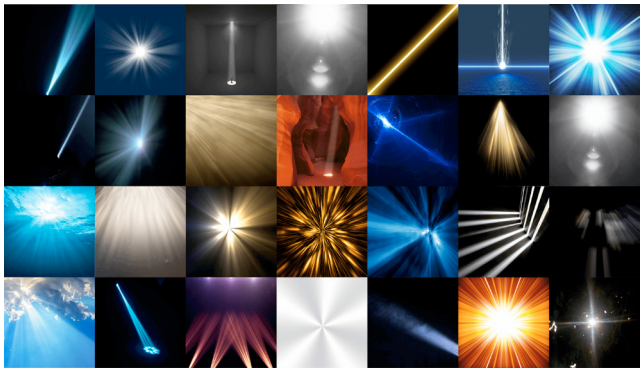

Fig. 1. An example of images collected for the dataset of "Ray of Light" lyrics line.

(4) We synchronized and sequenced the interpolation loops along the audio file to create the music video.

3.1 Image Collection

As a first step, we extracted the search queries from the lyrics and collected the images for the datasets. To create a semantic link between the song's lyrics and the visuals, we simply used each individual lyric line in the song as the initial search query. We then used these search queries to collect the images. We also augmented the search terms by using single words and idiomatic phrases from the lyrics and adding other lexical constituents (e.g., synonyms). For example, for the lyrics line, "I'm the first ray of light", we used the search queries "Ray of light" and "Beam of light". To further improve the search results, we appended the tags such as "photography", "art", or "painting" to the search queries. The images were sourced by using Bing Image Search API, which enabled us to collect the images pragmatically without scraping the hosting pages. As we expected that the pictures might point to links that do not contain a valid image, we also verified that each image is an actual image file. Table 1 shows selected examples of lyric lines along with the search queries used to collect the images as well as the number of pictures collected. Additionally, Fig. 1 shows an example of a subset of images collected for the lyrics line "Ray of Light".

Table 1. Lyrics Dataset Example

Lyrics	Search Queries	Number of Images
I'm the first ray of light	"Beam of light", "Ray of light"	411
I'm the moment of peace	"Calmness", "Peaceful moments", "Stillness", "Calm images"	1230
Starting inspiration	"Spiritual", "Inspiration photography", "Pictures of inspiration", "Abstract art inspiration"	1413
I'm the seed of change	"Change painting", "Seed painting", "Evolution painting"	770
Born in death	"Birth painting", "Creation artwork", "Death painting", "Birth abstract art"	1465

3.2 Image Clustering

While search engines provide a quick and accessible way to collect large datasets, many queries return non-

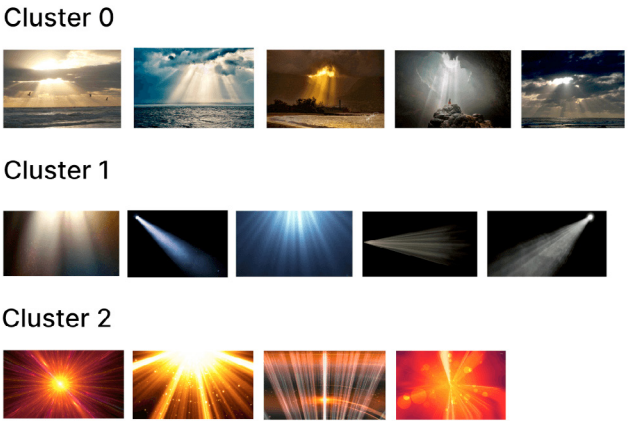

Fig. 2. An example of three clusters for the "I'm the first ray of light" lyrics line.

representative images or noise. Therefore, the artists should have the artistic freedom to select what visual representation they would like to associate with their music. Currently, many image search engines fall short when given an abstract concept as a keyword. For example, abstract keywords such as "calmness" will not provide representative images. Instead, searching for a related idea (e.g., a person taking a deep breath) would produce more accurate results. For example, *Initiation* is a song about existence and unfolding reality. Naturally, the song's lyrics are abstract, and it is challenging to translate them into representative semantic keywords.

As a way to tackle this problem, we grouped the images in k clusters based on visual similarity. Then, we selected and merged the sets of visually similar images. We used a pre-trained convolutional neural network (CNN) VGG16 (Simonyan and Zisserman (2014)) as a feature extractor by removing the final (e.g., prediction) layer to obtain a feature vector. Then, we clustered the images using K-means. We have experimented with a number of pre-trained models such as InceptionV3 (Szegedy et al. (2016)), Xception (Chollet (2017)), VGG19 (Simonyan and Zisserman (2014)) and ResNet50 (He et al. (2016)) which are considered to be state of the art for image recognition tasks. We experimented with different values as a number of clusters and selected $k = 10$ f as we noticed that this number of clusters yields the most optimal results. Finally, we handpicked the clusters that represent the concept of the lyrics. An example of the images collected with the search query "Ray of light" grouped into $k = 3$ clusters is shown in Fig. 2. At the end of the process, we merged the images collected from separate search queries that correspond to the same lyric line (such as "Ray of Light" and "Beam of Light") in one collection. An overview of the process is visualized in Figure 3.

As a next step, we manually reviewed the collected images and removed any noisy or otherwise outlier images. These were usually images with some artifacts, low resolution, or pictures that were non-representative of the concept.

3.3 Model Training

Generative adversarial networks (GANs) were first introduced by Goodfellow et al. (2014). Since their introduction, many researchers have expanded on using adversarial networks to create realistic samples. For our baseline, we have considered a number of GAN-based image generation models such as BigGANs (Brock et al. (2018)), BigBiGAN (Donahue and Simonyan (2019)), StyleGAN2 Karras et al. (2020b) and StyleGAN2-ADA (Karras et al. (2020a)) to generate the images. BigGANs generates images with relatively high resolution of up to 512 x 512 pixels. BigBiGAN, based on BigGANs, extends the model by adding an encoder module and modifying the discriminator. StyleGAN2-ADA is a state-of-the-art network that generates realistic images and provides results matching StyleGAN2 *"with an order of magnitude fewer images"* (Karras et al. (2020a)). Based on initial testing, we settled on StyleGAN2-ADA as it provides higher-quality results with significantly less training data. We did not consider using the newer version of the model, StyleGAN3 (Karras et al. (2021)), as it was released too recently. As a first step, we transformed the images to the format required by the model (e.g., the images need to be square-shaped and be of the power-of-two dimensions). Therefore, we resized the images to 1024 x 1024 pixels and converted the images to `.png` image format. Next, we converted the images to `TFRecord` type, which is a format for storing a sequence of binary records. We then made use of the High-Performance Computing (HPC) cluster (Varrette et al. (2014)) of the University of Luxembourg to allocate the training tasks to the GPU nodes. By using the HPC facilities, we have significantly improved the training time compared to training the model on the local machine or cloud computing platforms (e.g., Google Colab). In total, we have trained 15 StyleGAN2-ADA models. Each model was trained on about 10000 iterations (*kimg*). The training of each model took approximately three days on a node with 4 Tesla V100 GPUs. As an example, the training progress for the lyrics line "I'm the first ray of light" is demonstrated in Fig. 4.

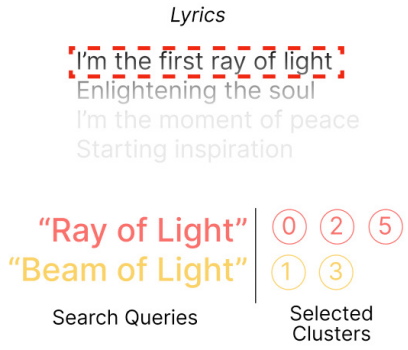

Fig. 3. An illustration of the process of collecting the images, selecting clusters and merging search queries.

3.4 Generating Interpolations

Once the training process was complete, we used the latest model checkpoints to generate interpolation loops. By interpolating between random images (e.g., seeds) in the latent space, we created a seamless transition between the various images generated by the model. Each interpolation loop is 15 seconds long and features nine different images

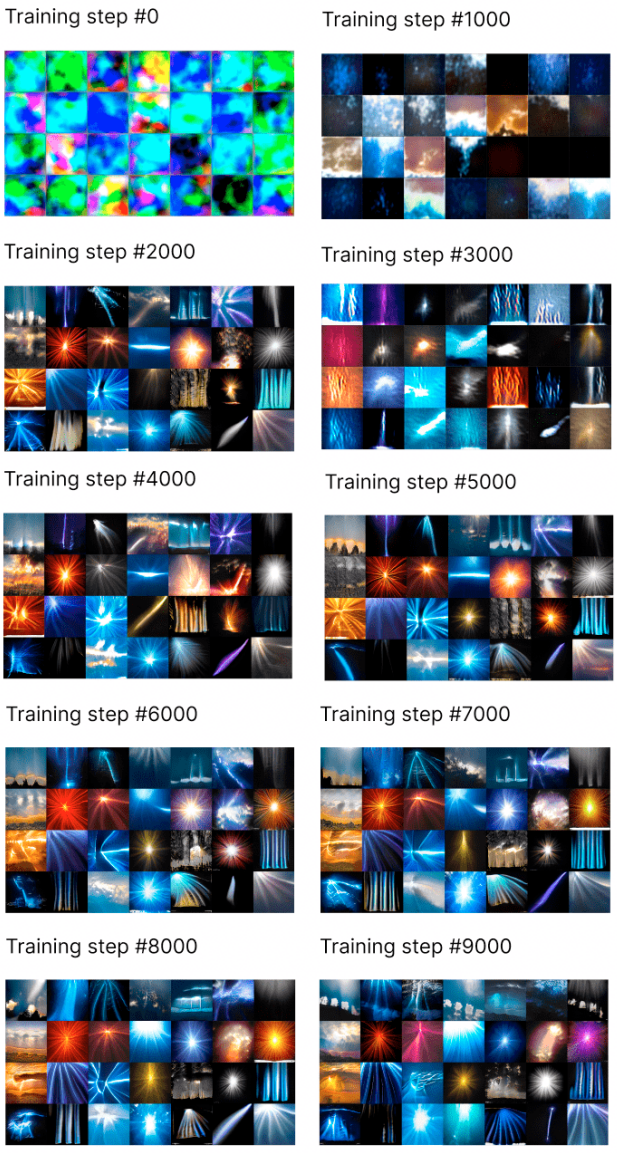

Fig. 4. The training process over 9000 training iterations for the "I'm the first ray of light" dataset. The values are in *kimg*.

(e.g., seeds). Furthermore, we selected the same starting and ending image, which creates an illusion of an infinite loop while on playback. As a resulting film, we made an interpolation loop for each of the 15 lyric lines and combined them in one final cinematic film by using Final Cut Pro X. The resulting lyric video is 3 minutes and 45 seconds long (Gareev and Glassl (2022)).

3.5 Results

As a result, we have collected 15 datasets with on average 850 images in each. We grouped the images in a total of 15 datasets which represent each of the 15 lyric lines in the song. The overall size of all datasets combined is approximately 7 GB. From all of the images collected, 12750 were used as training data. The combined training time of 15 StyleGAN2 models on the HPC facilities was approximately 45 full days or 1080 hours. The resulting film has a resolution of 1024x1024 pixels and a frame rate of 60 frames per second. We have open-sourced the code for data collection, image clustering, and model training on Github so that the details of the technical implementation can be further explored (Gareev (2022)).

4. FUTURE WORK

While we are in the early stage of research, we recognize that currently, our approach has many limitations in terms of usability (e.g., creators without any technical knowledge are unable to generate a lyric video). As a future work, we would like to develop a user interface (UI) that allows creators in creative practice to input the lyrics and then interactively create, arrange and export AI-generated music videos. This includes automatically augmenting and refining search terms and providing an option to select the images to use as training data for the model. To arrive at the design, we will conduct user studies with various users at all stages of the development process. User studies will provide valuable insights into human-AI interaction as well as the design, implementation, and evaluation of AI-based generative interfaces. This will support the generalization to other songs and produce an end-to-end solution.

Additionally, we plan to experiment with more sources of images for the model (e.g., stock photography resources such as Unsplash (Wikipedia (2022b))). Specifically, we are interested in developing a search algorithm using natural language descriptions. For example, OpenAI's CLIP neural network (Radford et al. (2021)) can transform images and text pairs into the joint latent space. Then, they can be compared using a similarity measure. It would be interesting to pre-train images with CLIP and then use the pre-computed feature vectors of the images to find matches to a natural language search query. We expect that the interface based on the natural language search will provide better efficiency than the keyword search as the keyword search has limitations on retrieving images that represent abstract concepts.

Currently, our approach can generate static lyric videos. They do not react to the song's structure (e.g., timbre, tempo, etc.). When StyleGAN2-ADA generates images, it uses a latent vector containing 512 numbers. It would be worthwhile to automatically alter this input vector using various audio variables (e.g., amplitude, frequency, etc.). This will affect the output images and produce a lyric video that is synced with music's dynamics, rhythm, and structure.

Finally, the work on generative models involves significant and broad social impacts. In the future, we would like to analyze how AI-based generative technology can potentially have a bias in the model outputs as well as the longer-term ethical challenges.

5. DISCUSSION AND CONCLUSION

We have presented a simple approach for creating AI-generated lyric videos. By collecting the training data for each lyric line and using StyleGAN2-ADA to generate the images, we generated representative images for each lyric line in the song. We then seamlessly interpolated between these images to create an aesthetically pleasing video

that represents the lyrics' visual narrative. The resulting lyric video (Gareev and Glassl (2022)) remains abstract and leaves a considerable headroom for interpretation by the spectator. As the lyrics' concepts are abstract and supposed to stimulate the listener's imagination, this is an appropriate outcome. As lyric lines represented a wide range of semantic notions, we had to derive search queries from the lyrics and augment them with synonyms and other lexical constituents to narrow down the specific desired concepts. As we had significant artistic freedom to influence the model outcome, the illustration of AI consciousness depicted in these images remains within the artistic limits. We hope that this experiment will enable musicians to explore new forms of creative expression, where artistic expression via lyrics can produce an AI-generated lyric video. We recognize that this work is still in an early stage of research, and there are still several constraints that render this approach unusable to many non-experienced users. First, the training with StyleGAN2-ADA is computationally expensive and time-consuming. Second, the users must possess the technical expertise to make use of the technique.

Generative AI is often viewed as creative due to its ability to produce work indistinguishable from the art created by humans. The visual arts created by AI-based generative systems are often perceived as surprising, interesting, or visually appealing to many people. We presented the lyric video at an AI and art exhibition, and we were surprised by how the AI-created artworks were perceived. Many visitors mentioned that the lyric video has a high degree of expression and provides a pleasing aesthetic experience. GAN-based approaches have been widely used in the creative landscape to generate realistic artworks in the last few years. Recent advances in the development of multimodal generative models (e.g., text-to-image models) could have an even more significant impact on the production and consumption of art. Translating data from multiple modalities into a joint semantic space is a promising approach for creative exploration as the concept of multimodality is intrinsic to many forms of art. The AI-generated example presented in this paper demonstrates the increasingly sophisticated potential of AI to produce art that provides aesthetic value. With recent breakthroughs in multimodal research, AI systems will continue to become more relevant in the production and analysis of art. They will continue to improve to produce convincing forms of art that imitate human-created works of art. However, this does not imply that these interfaces should be completely autonomous. Therefore, the research on human attitudes and perceptions toward AI art is a novel area of research that requires more examination and discussion as it has implications on human creation and understanding of art.

ACKNOWLEDGEMENTS

We would like to acknowledge and thank our supervisors from the University of Luxembourg, who provided insight and expertise that greatly assisted the research. The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al. (2014)).

REFERENCES

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

Cetinic, E. and She, J. (2022). Understanding and creating art with ai: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2), 1–22.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Donahue, J. and Simonyan, K. (2019). Large scale adversarial representation learning. *Advances in Neural Information Processing Systems*, 32.

Frosst, N., Kereliuk, J., and Kid, G. (2019). Text conditional lyric video generation. In *chap. Machine Learning for Creativity and Design Workshop*.

Gareev, D. (2022). Stylegan2 repository. URL `https://github.com/lowlypalace/StyleGAN2`.

Gareev, D. and Glassl, O. (2022). Initiation lyric video. URL `https://youtu.be/_JVv8CtXVKA`.

Glassl, O. and Savo, M. (2019). Initiation. URL `https://distrokid.com/hyperfollow/thalamusproject/initiation`.

Glassl, O. and Savo, M. (2022). Thalamusproject. URL `https://www.thalamusproject.com/`.

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

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020a). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33, 12104–12114.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020b). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Torrico, T.J. and Munakomi, S. (2019). Neuroanatomy, thalamus.

Varrette, S., Bouvry, P., Cartiaux, H., and Georgatos, F. (2014). Management of an academic hpc cluster: The ul experience. In *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*, 959–967. IEEE, Bologna, Italy.

Ward, L.M. (2013). The thalamus: gateway to the mind. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(6), 609–622.

Wikipedia (2022a). Esch 2022. URL `https://lb.wikipedia.org/wiki/Esch_2022`.

Wikipedia (2022b). Unsplash. URL `https://en.wikipedia.org/wiki/Unsplash`.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.