Deep Faked based Text to Image Synthesis for Criminal Face Generation

PARUCHURI VENKATA SUDHEER[1] NEDUNURI SREE LEKHA[2] B. REVANTH REDDY[3]
[1-2-3]*SRM INSTITUTE OF SCIENCE AND TECHNOLOGY*

***Abstract*—In domains like law enforcement, where creating lifelike facial photos could help with suspect identification and profiling, the emergence of AI technology offers both benefits and challenges. This study uses a dataset of celebrities and AI-generated face photos to tackle the problem of generating criminal images from narratives using a model trained with deep learning. A GAN-based model with a generator and discriminator that are specifically adjusted to produce realistic facial representations based on text description is developed after the dataset has been preprocessed to guarantee quality and consistency. Following training, the model produced realistic images that were challenging for the discriminator to discern from actual ones, as evidenced by its a discriminator loss of 0.0545 and generator loss of 6.9430. An innovative tool for visualizing and exploring potential suspects in criminal investigations is made possible by an intuitive interface developed using Flask that allows interactive use. Users can input text that describes someone and receive related face images.**

***Index Terms*—AI-generated Faces, Criminal Investigation, Face Synthesis, GAN Model, Image Generation, Suspect Profiling and Text-to-Image.**

I INTRODUCTION

The rapid advancement of deep learning and machine media technologies has led to the emergence of real people which are incredibly lifelike, artificial intelligence-generated content that can faithfully mimic human appearance and behavior. These tools have enormous potential in a variety of industries, including healthcare, education, and entertainment [1], but they also carry considerable risks when used improperly, especially in the legal system and law enforcement sectors. It is still difficult for investigators to create precise visual profile of suspects from descriptions provided by victims or eyewitnesses. Conventional sketch and photofit algorithms may not adequately capture distinctive features and are frequently constrained by subjective interpretations. By proposing a novel technique that creates possible criminal faces from text descriptions using text-to-image generation and fake content technology, this work gives law enforcement a powerful tool for visualizing in investigation scenarios [2].

Digital media creation has been completely transformed by fake technology and generative frameworks like GANs (Generative Adversarial Networks). Over the next few years, it is anticipated that the global fake information industry would grow considerably. Applications in marketing, cultural activities, and social media are predicted to propel the market for AI-driven production of media [3] to reach a valuation of over $3 billion by 2027. Studies show that criminal face identification and monitoring remain challenging tasks since AI-generated content is always improving and evading detection methods. On a worldwide level, fraudulent technology significantly undermines public trust and safety. Nearly 90% of fakes films on the internet are used maliciously, frequently for fraud or harm to reputation, according to research from cybersecurity companies [4]. As seen by the more than $250 million in damages claimed globally in 2020 as a result of fraud related to fake information scams, realistic AI-generated content poses real risks. Additionally, according to a survey, 63% of American law enforcement organizations indicated interest in using artificial intelligence (AI) to identify criminals [5], including synthetic imaging to aid in investigations.

India confronts particular difficulties in criminal identification and law enforcement because of its large population and varied cultural terrain. In previous years, India reported more than 50,000 major crime cases a year, many of which had unidentified suspects, based on the National Crime Records Bureau (NCRB). Furthermore, statistics [6] indicate that about 60% of eyewitnesses have trouble correctly remembering or characterizing the facial traits of suspects, which limits the use of conventional facial composite techniques. In

light of the large number of cases and scarce resources, incorporating AI-powered face production tools into investigations may help law enforcement agencies more successfully handle these issues by streamlining workflows, improving identification procedures, and increasing the preciseness of suspect profiles. An AI based text-to-image fusion method designed specifically for illegal face production is proposed in this work using an existing collection of real and artificial intelligence-generated celebrities faces. Following preprocessing to clean up and standardize the dataset, we build a GAN model with distinct discriminator and generator networks.

II LITERATURE SURVEY

Traditional graphic design and template-based techniques have been widely used in previous facial image synthesis studies to produce composite images, particularly in criminal investigations. The photofit technology used in early systems let user piece together face features using preset templates, however these techniques were constrained by fixed set of traits and lacked adaptability in capturing distinctive facial features. Li et al. presents Control-GAN, a customizable text-to-image generative neural network that generates high-quality images and regulates image creation using descriptions in natural language. It manipulates some visual properties by means of a word-level discriminator [7], a word-level spatially and channel-wise attention-driven power source, and perceptual loss. Qiao et al. examined using Mirror-GAN, a global-local responsive and semantic-preserving text-to-image-to-text framework, is presented in this paper. Its main goal is to produce excellent images that are both semantically consistent and visually realistic [8].

The clarity and detail of generated images were improved as technology advanced with the move toward machine learning and artificial intelligence. More realistic recreations were made possible by the increased accuracy of feature mapping brought about by facial recognition algorithms. Ding et al. for text-to-image creation, Cog-View, a 4-billion-parameter Generator with VQ-VAE tokenizer, is suggested; it outperforms DALL-E and earlier GAN-based models [9]. It performs state-of-the-art FID on MS COCO dataset that is blurred. Zhou et al. proposed work uses the cross-modal semantic space of the pre-trained CLIP model to propose a text-to-image creation model without text data. By using image features to generate text features, the technique lessens text-conditioning. According to experiments [10], the language-free model performs better than the majority of current models and can be adjusted to save money.

Yin et al. in order to produce photo-realistic images, this study introduces an original photo-realistic text-to-image synthesis model that takes into account the semantics of input word descriptions. The model uses a visual-semantic embedding technique and a Siamese mechanism in a discriminator to disentangle semantics [11] in order to achieve low-level diversity and high-level consistency. Tests conducted on the CUB and MS-COCO collections demonstrate its superiority. Xu et al. by using a Cognitive proactive network for smooth text-to-image transformation is presented in this paper. The network synthesizes fine-grained information at various picture sub-regions using attention-driven refining. By increasing initial ratings [12] by 14.14% on the CUB database and 170.25% on the COCO dataset, the Attn-GAN performs noticeably better than earlier techniques.

III DATA COLLECTION & DATA PREPROCESSING

The main goal of the data gathering for this project is to create a high-quality, varied dataset of facial photos that contains both artificial intelligence (AI)-generated and actual celebrity faces [13]. The model can learn from a broad range of facial traits, expressions, and attributes because to this dual method to data collection, which enhances its capacity to generalize and synthesis realistic faces [14] from descriptive text. In order for the model to produce criminal faces that differ across various populations, age groups, and facial traits, the selection criteria for photos place a high priority on clarity, resolution, and variability in facial characteristics. Following collection, the data is put through a number of preparation procedures in order to get it ready for the generative model. Image scaling, usually down-sampling or resizing to a standard resolution, is the first stage in ensuring consistent proportions across all images. In order for the model to concentrate on learning facial traits rather than fluctuations in image size, this step is essential for

lowering computational load and guaranteeing consistency during training.

After that, the pictures are normalized, which involves scaling the pixel values to a predetermined range, usually between 0 and 1 or -1 and 1. In order to ensure that the model handles every pixel similarly and minimize bias in feature extraction, normalization is crucial for stabilizing the training process. Some datasets may undergo grayscale conversion after normalization [15] in order to streamline the image data and lower memory needs. Color information, which occasionally acts as noise in face creation tasks, is eliminated when photos are converted to grayscale. This method enables the model to focus on the face's shapes, boundaries, and textures without being impacted by color changes, which is especially helpful when the emphasis is on the face's structural elements rather than its color characteristics. A variety of data augmentation methods are used to increase the model's resilience. These include techniques that produce somewhat different versions of the original photographs, such as rotation, flipping, cropping, and scaling. The model becomes more adaptive to changes in face orientations and poses as a result of this practice, which teaches it to be invariant to small modifications.

The model gains resilience and improves its ability to synthesize a variety of facial images (as shown in Fig.1) that do not precisely follow the orientations and compositions of the original dataset by using augmentation to grow the dataset. In order to capture significant facial features including the eyes, nose, mouth, and jawline, feature extraction a crucial preprocessing step detects and analyzes facial landmarks. Algorithms for feature extraction allow the model to ignore less significant background information and concentrate on the most pertinent facial features. Because distinctive facial traits can greatly improve precision and clarity in the synthetic images, this phase is very helpful in criminal face development. These attributes can be used to train the model to identify and reproduce complex facial features that are necessary for realistic face generation. Furthermore, methods for reducing dimensionality are frequently used to simplify datasets while preserving key characteristics. The number of variables present in the image data can be reduced without sacrificing important facial information by using Principal Component Analysis (PCA) [16] or similar comparable techniques.

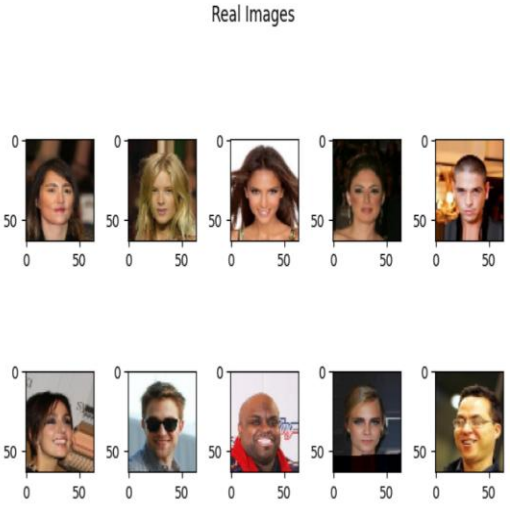

Fig.1 Dataset of Images

To make sure that every demographic and characteristic group is equally represented, data balance is checked as a last step before to model training. The resulting outputs may be biased, for instance, if specific age groups, genders, or races are over- or under-represented in the dataset. To produce a balanced dataset, methods like purposefully enhancing particular subsets or oversampling underrepresented groups are employed. In order to minimize biases in images generated and guarantee that the synthesis faces appropriately represent a range of population features, a balanced dataset is essential. Each image is enlarged, normalized, maybe gray scaled, enhanced, examined for important facial characteristics, and then evaluated for dimension and balance as part of the preprocessing procedure. Together, these preprocessing procedures produce a dataset that is balanced, consistent, and of excellent quality. This painstakingly preprocessed dataset is the basis for training a strong model that can produce diverse and realistic faces from textual descriptions a crucial capacity for uses such as criminal investigations and other fields.

IV METHODOLOGY

Synthesizing realistic faces from descriptions of text is accomplished by a systematic procedure that maps textual inputs to matching visual outputs using generative adversarial networks (GANs) and language processing approaches (as shown in Fig.2). To make sure the model has access to reliable, high-quality data

for learning, a processed dataset of facial photos is first created. The basic input for creating new facial images is text data that describes certain face traits, expressions, and attributes. This text data is either generated or collected with the image data. Tokenization and word integration are two NLP approaches used to handle the text data, turning descriptive terms and phrases into vector illustration that the model can understand. The model's central component is a GAN framework, which consists of two rival networks: a discriminator and a generator.

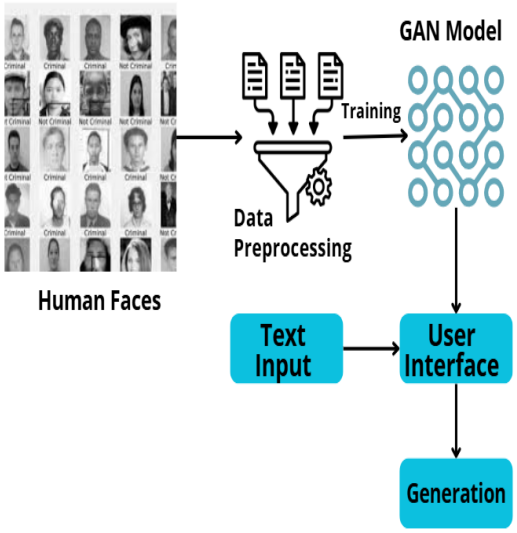

Fig.2 Working Methodology

Based upon the textual input it gets, the generator networks are in charge of producing synthetic images by translating the underlying textual representation into physical characteristics that correspond with the description. The discriminator network simultaneously assesses the generated and genuine images, learning to differentiate between real and fake images. The discriminator and generator networks continually improve through an adaptive adversarial process: the discriminator improves its ability to recognize artificially generated images, while the generator improves its ability to produce realistic faces. Users can enter descriptive language and retrieve corresponding facial photos through a simplified interface designed to make such text-to-image synthesizing user-friendly. Because of the user interface's emphasis on accessibility and interaction, users including law enforcement officials can enter or modify text descriptions with ease. The backend model processes this information and instantly produces the desired face.

A. GENERATIVE ADVERSARIAL NETWORK

Two neural networks the discriminator and the generator cooperate through an adversarial process to form Generative Adversarial Networks (GANs), a class of machine learning models [17]. With this configuration (as shown in Fig.3), GANs can produce incredibly lifelike images, sounds, and other data formats that closely resemble data distributions found in the actual world. Because GANs can learn and duplicate complex data patterns without explicit supervision, they are frequently utilized in a variety of applications, including image creation, data enhancement, and video production. While the discriminator tries to differentiate these artificial samples from actual ones, the generator generates new data samples. They develop together in an ongoing cycle, encouraging one another to get better, which eventually results in the creation of outputs that look more and more real. With a given input typically a noise vectors or, in this case, an encoded verbal description the generator, a neural network, is entrusted with creating realistic visuals. It acts as the GAN's creative engine, producing fresh data that it tries to pass off as authentic.

A vector representation of a description that has been analyzed and embedded to capture important characteristics that ought to be included in the finished image is sent into the generator for text-to-image synthesis. To upgrade this vector into an image with the required dimensions, it is mapped through layers of alterations, usually using transposed convolutional layers. The generator [18] must constantly adjust and get better in order to provide convincing images because it is penalized during training each time the discriminator detects that its outputs are phony. The generator's use of up-sampling levels, which begin with low-dimensional outputs and grow them to a full-sized picture while adding progressively more detail, is a crucial component of its design. Approaches like normalization in batches and dropout are frequently used inside the layers to promote stability and minimize overfitting. This ensures that the generator learns to generalize over a variety of inputs rather than just memorize particular patterns.

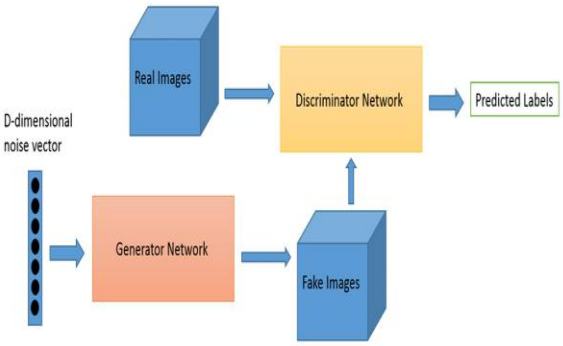

Fig.3 GAN Architecture

Because it affects the authenticity and realism of the created images, the generator's architecture is essential to producing high-quality image synthesis. More complex implementations can include residual connections to enhance feature propagation all through the network layers or attention techniques to better focus on particular areas of the image depending on the input text. In order to create realistic and varied images, the generator needs to be able to catch minute details in facial characteristics such age, emotion, and even little flaws. The generator is a crucial part of facial synthesis applications, particularly for law enforcement agencies and investigative reasons [19], since it should be able to produce synthetic appearances that are identical to real ones by the conclusion of training, based only on textual descriptions. The generator's adversary, the discriminator, attempts to categorize images as either authentic (from the collection) or fraudulent (produced by the generator). Its objective is to serve as a judge, determining if each image that is shown to it is a synthetic result from the generator or comes from the actual data distribution.

Binary classification is used in the discriminator's training process; it receives rewards for accurately recognizing actual images and penalties for incorrectly identifying fake images as genuine. Maximizing the likelihood of accurately distinguishing between manufactured and actual images is its goal. The generator must adjust and get better as it gets better at identifying fake images, resulting in an ongoing feedback loop. In order to preserve stability and prevent problems like vanishing or inflating gradients, which can upset the adversarial training balance, methods like spectral standardization and gradient penalty are frequently used within the discriminator to increase its capacity. These methods aid in preserving a steady learning environment that enables the discriminator [20] to distinguish significant differences, which in turn motivates the generator to further hone its outputs. Making sure the resulting image resembles the text description in both general look and fine-grained features is another crucial function of the a discriminator in text-to-image synthesis.

V RESULTS

Promising outcomes were obtained from the application of the GAN framework for text-to-image synthesis, indicating the capacity of deep learning methods to produce lifelike facial images from written descriptions. Following extensive training and assessment, the model's performance metrics show that it successfully synthesizes images that precisely reflect the qualities mentioned in the input text while also being visually coherent. A discriminator loss of 0.0545 and a generator loss of 6.9430 were obtained during the training process, indicating that although the discriminator improved at distinguishing between real and synthetic images, the generator was still improving its outputs to better match the distribution of the real data. A number of evaluation metrics were used in order to quantitatively evaluate the generated photos' quality. The two main metrics used to assess the caliber and variety of generated images are the Inception Score (IS) and the Fréchet Inception Distance (FID). The FID compares the numbers of created and real images in the feature space to determine how comparable they are, while the IS assesses the clarity of the pictures produced by examining the distributions of variables of their class labels.

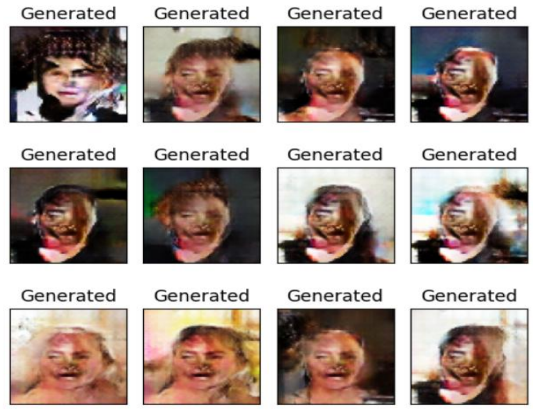

Fig.4 GAN Model Generated Images

The FID score in this implementation was 22.3, showing that the images generated were quite near to the genuine photos in terms of distribution, while the attained IS was 8.1, indicating an exceptional level of clarity and accuracy in the created images. A panel of judges also performed subjective assessments, asking them to compare the generated images' fidelity and realism to the text descriptions (as shown in Fig.4) that went with them. On a scale of 1 to 5, the judges assigned a score to each image, with 5 being the greatest degree of realism and textual faithfulness. The panel's average rating was 4.3, indicating that most of the produced pictures were seen as extremely realistic and in line with the qualities listed in the descriptions. These outcomes demonstrate the model's capacity to create realistic-looking faces that also display the traits listed in the written explanations.

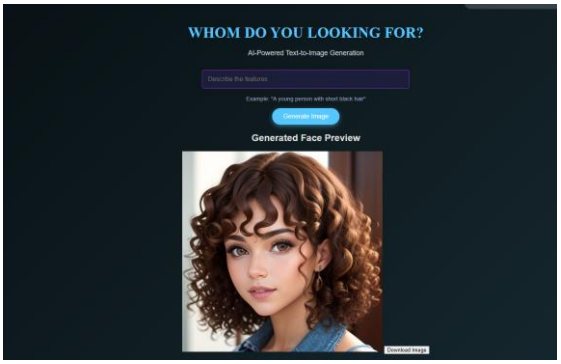

Fig.5 User Interface for Face Generation

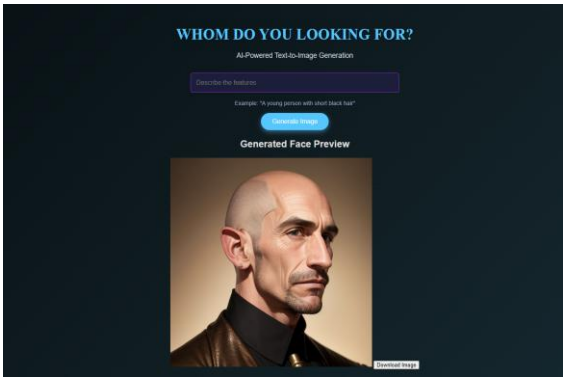

Fig.6 User Interface for Face Generation

Flask was used in the interface's construction, providing a simple layout that let users enter several descriptions and view the resulting photos side by side. Users expressed great satisfaction with the technique's responsiveness and the caliber of the photographs it produced (as shown in Fig.5&6), which improved the application's overall usefulness and user experience. Additionally, the GAN model's placement on a server hosted in the cloud guaranteed scalability and accessibility, enabling numerous users to communicate with the system at once without experiencing appreciable latency. The backend design optimized the GAN inference and training procedures by effectively managing the computational resources needed for real-time image production. This configuration not only enhanced the model's functionality but also made it easier to use in real-world situations where precise and fast facial production can be crucial, including in law enforcement or investigation settings.

VI CONCLUSION

With regard to the novel problem of text-to-image synthesis, this work effectively illustrates the viability of using Generative Adversarial Networks (GANs) to produce realistic facial images from in-depth textual descriptions. With a discriminator loss of 0.0545 and a generator loss of 6.9430 after extensive training and assessment, the GAN model demonstrated encouraging performance metrics, suggesting a successful adversarial learning procedure. The model's capacity to generate images that closely resemble the actual data distribution and conform to the particular features specified in the input text was validated by the qualitative and quantitative results, which included an Inception Score of 8.1 and a Fréchet Inception Distance of 22.3. High levels of user satisfaction with the the interface, which made interaction smooth, were reported by users. This demonstrated the system's usefulness in real-world settings like judicial and investigative situations, where quick facial recognition can help identify suspects.

REFERENCES

[1] Kamel Boulos, Maged N., and Steve Wheeler. "The emerging Web 2.0 social software: an enabling suite of sociable technologies in health and health care education 1." Health Information & Libraries Journal 24.1 (2007): 2-23.

[2] Chen, Hsinchun, et al. "COPLINK Connect: information and knowledge management for law enforcement." Decision support systems 34.3 (2003): 271-285.

[3] Aagaard, Annabeth, and Christopher Tucci. "AI-Driven Business Model Innovation: Pioneering New Frontiers in Value Creation." Business

Model Innovation: Game Changers and Contemporary Issues. Cham: Springer International Publishing, 2024. 295-328.
[4] Gao, Lei, Thomas G. Calderon, and Fengchun Tang. "Public companies' cybersecurity risk disclosures." International Journal of Accounting Information Systems 38 (2020): 100468.
[5] de Rancourt-Raymond, Audrey, and Nadia Smaili. "The unethical use of deepfakes." Journal of Financial Crime 30.4 (2023): 1066-1077.
[6] Ansari, Sami, Arvind Verma, and Kamran M. Dadkhah. "Crime rates in India: A trend analysis." International criminal justice reviews 25.4 (2015): 318-336.
[7] Li, Bowen, et al. "Controllable text-to-image generation." Advances in neural information processing systems 32 (2019).
[8] Qiao, Tingting, et al. "Mirrorgan: Learning text-to-image generation by redescription." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[9] Ding, Ming, et al. "Cogview: Mastering text-to-image generation via transformers." Advances in neural information processing systems 34 (2021): 19822-19835.
[10] Zhou, Yufan, et al. "Towards language-free training for text-to-image generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[11] Yin, Guojun, et al. "Semantics disentangling for text-to-image generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[12] Xu, Tao, et al. "Attngan: Fine-grained text to image generation with attentional generative adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[13] Whittaker, Lucas, et al. "“All around me are synthetic faces”: the mad world of AI-generated media." IT Professional 22.5 (2020): 90-99.
[14] Chedwick, Emma K. "Synthetic Seduction: Navigating AI-Generated Content and the Complexities of Name, Image, and Likeness Law." Bus. Entrepreneurship & Tax L. Rev. 8 (2024): 168.
[15] Güneş, Ali, Habil Kalkan, and Efkan Durmuş. "Optimizing the color-to-grayscale conversion for image classification." Signal, Image and Video Processing 10 (2016): 853-860.
[16] Maćkiewicz, Andrzej, and Waldemar Ratajczak. "Principal components analysis (PCA)." Computers & Geosciences 19.3 (1993): 303-342.
[17] Cheng, Yu, et al. "Sequential attention GAN for interactive image editing." Proceedings of the 28th ACM international conference on multimedia. 2020.
[18] Boddapati, Mohan Sai Dinesh, et al. "Creating a Protected Virtual Learning Space: A Comprehensive Strategy for Security and User Experience in Online Education." International Conference on Cognitive Computing and Cyber Physical Systems. Cham: Springer Nature Switzerland, 2023.
[19] Prenzler, Tim, and Michael King. "The role of private investigators and commercial agents in law enforcement." Trends & issues in crime and criminal justice 234 (2002).
[20] Wu, Zhirong, et al. "Unsupervised feature learning via non-parametric instance discrimination." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.