



DL-based Generation of facial portraits from diverse data sources

Mohammed BERRAHAL and Mostafa AZIZI

MATSI Research Lab, ESTO, Mohammed First University, Oujda, Morocco

Received 3 Jun. 2022, Revised 19 Dec. 2022, Accepted 6 Feb. 2023, Published 16 Apr. 2023

Abstract: Artificial intelligence has a tremendous potential to reinforce and revolutionize law enforcement. By automating the tedious and time-consuming tasks of data collection and analysis, AI can help police departments become more effective. In this paper, we provide an overview of possible deep learning models that can be utilized by law enforcement to reconstruct the suspect's face, through generating real and sketch facial portraits. To accomplish this, we collect four types of data from the crime scene, handwritten text and audio from an officer's note, images from a smartphone, and video from a surveillance camera. Then, employ two pre-trained models: ABM-CNN for attribute multi-label classification and Google Speech API for speech-to-text conversion. In addition, we train three other models, the first of which, a handwritten model trained on the IAM Handwriting dataset, reads and digitizes handwritten notes with an accuracy of 76%, outperforming state-of-the-art results. Second, we train YoloV5 with the Wider Face dataset to detect one or multiple faces on images or videos with an average precision of 93%, a recall of 90%, and a precision of 88%. In the third model, we adapt the Zero-Shot Text-to-Image Generation technology to generate real faces and sketch. Our resulting model outperforms existing models from literature regards to high-quality, and the training reach a loss of 37.4%.

Keywords: Portrait Generation, Text-to-image Synthesis, Deep Learning, Handwritten OCR, YOLOv5, Face Detection.

1. INTRODUCTION

Face-matching, facial composites are a common tool used by detectives in complex cases, police officers attempt to find out the person involved, by using all possible information obtained from a crime scene, provided by multiple sources, to find the suspect's identity or to reconstruct his face [1][2]. The resulting image, the sketch, is circulated among members of the police force or in media outlets such as television, newspapers, and social media to identify the suspect, locate him, or provide a lead in the investigation.

Due to the increase in criminal activities, it is necessary to keep up with the technological advancements, to be able to reduce the scope of an investigation and increase the efficiency. The use of deep learning and computer-aided machine vision is an essential tool to help us reconstruct a real portrait image [3].

There are four main sources of information's capable of identifying face suspects, the first one, is the notes taken by a police officer, by asking questions on eyewitnesses [4], to extract facial attributes of the offender, and the second is videos from a close surveillance camera, the third is by taking a photo from people close the crime scene, the

last source is an audio description from an eyewitness. To generate a complete portrait of the offender, we must have at least, one source of information that we already mentioned.

Identifying and generating a face requires training a system on the facial characteristics that make each person unique. The smile, the presence of a beard and black hair, the wearing of scarf or a necklace, and so on are all examples of facial features that are intuitive semantic aspects that characterize the visual properties of face pictures that humans can perceive [5][6].

In our work, we are going to focus on generating a complete portrait, based on all possible information gathered from the crime scene [7]. This information is collected from different sources and has different formats: handwritten notes, audio, images, and video as shown in Figure 1. All data has to pass through a deep learning model to be transformed on the clear text description, we use two pre-trained models, one for image attribute classification model, named Augmented Binary Multi-label CNN (ABM-CNN) of our previous work [8], the second to transform audio to text, taking benefits from the Google Speech API for speech-to-text [9]. As for the other data type, we train two

models, the first one for handwritten notes to transform it into digital text, the second to detect faces on images, and videos.

The output text description from the previous models will be the input of the final model which will generate facial images, real and sketch. Finally, to overcome the lack of a facial dataset that regroups text description, real images, and sketched images, we train the StyleGAN model that can transform high-definition real images into a sketch [10].

The remainder of this paper is structured as follows: In the second section, we discuss the background and recent related works. The third section describes our research methodology. Before concluding, the fourth section presents our models' implementation and discussions.

2. BACKGROUND AND RELATED WORKS

In this section, we give a general vision of the deep learning models that we will use in our approach, and we go through related works dealing with models of classification, estimation, and generation.

A. Classification and estimation models

- Handwritten Text Recognition (HTR): also known as text character recognition, is a type of computer technology in which optically processed letters are recognized and interpreted [11]. The fundamental concept behind HTR is to turn any handwritten or printed text into data files that can be manipulated and read by machines. To digitize handwritten text, we divide the giving sentence in word and focus our model to predicting every word separately [12].
- Facial Attributes Classification (FAC): The most important aspect of FAC is its ability to predict several facial characteristics, on the given image or face portrait. There are two basic approaches to face attribute classification: local and global. Training a classifier for diverse qualities using landmarks is the main emphasis of local techniques. To extract a feature representation that is not dependent on landmark locations, global approaches focus on analysing the entire facial image [13].
- Face detection (Yolov5): YOLOv5 (You Only Look Once) is an object detection method that employs deep learning. To recognize objects in a picture, it uses a convolutional neural network that uses only one type of neural network. The method has a high degree of accuracy and may be used in real-time to identify things. In comparison to previous object detection algorithms, YOLOv5 has several advantages. For instance, it is capable of accurately detecting a large variety of items; it can also accomplish this detection in real-time, making it ideal for use in self-driving automobiles for example [14].
- Automatic Speech Recognition (ASR): is a method

for turning spoken words into text. ASR's performance has seen a significant boost thanks to the advent of deep learning. So many companies are already adopting ASR in their business strategies. The main work of ASR technologies relay on evaluating the vocal recordings from a speaker who reads text or a series of distinct words and relate them to the collection of texts.

Automatic voice recognition may be approached using Neural Networks, if their performance is enough. Short-term units like isolated words and phonemes were first categorized by the networks, which had a restricted skill set. Over time, as the complexity of neural networks like LSTM networks increased, the performance of these networks also improved [15].

B. Text-to-image synthesis models

In recent years, the Generative Adversarial Network (GAN) has emerged as among the most creative and productive deep learning approaches. GANs are neural networks that combine a generator network with a discriminator network. For starters, feeding a noise vector into a deep neural network with numerous convolutional layers to produce a picture, which then feeds an image generation neural network with multiple convolutional layers to generate an image. This image is then sent to a discriminator, which must determine whether it originates either from the generator or the actual dataset (the photos used for training). Learned signals from this discriminator network will be propagated throughout the model pipeline, which will eventually be able to generate images that appear extremely similar to the dataset [16]. This novelty led to several other areas which were impossible before, one of them is text-to-image synthesis.

H. Zhang et al have developed a system, named Stacked Generative Adversarial Networks (Stack-GAN). Stack-GAN is based on two stages of training; the first stage generates low-resolution images using rudimentary shapes and colors, and the second stage generates realistic images of high quality using the results of the first stage and text descriptions as inputs. The framework then employs a conditioning augmentation strategy to degrade the results of both stages [17][18].

Tao Xu et al present in their paper an Attentional Generative Adversarial Network (Attn-GAN). There are two primary parts in this framework: As a first step, the attentional generative network uses text to generate pictures with a poor resolution combined with the relevant text vectors to create new visuals. Secondly, the Deep Attentional Multimodal Similarity Model (DAMSM) provides an additional fine-grained picture-text match to compensate the training loss [19].

Ming-Tao et al develop a model named "Deep Fusion Generative Adversarial Networks (DF-GAN)". It is a framework that uses three significant advancements to

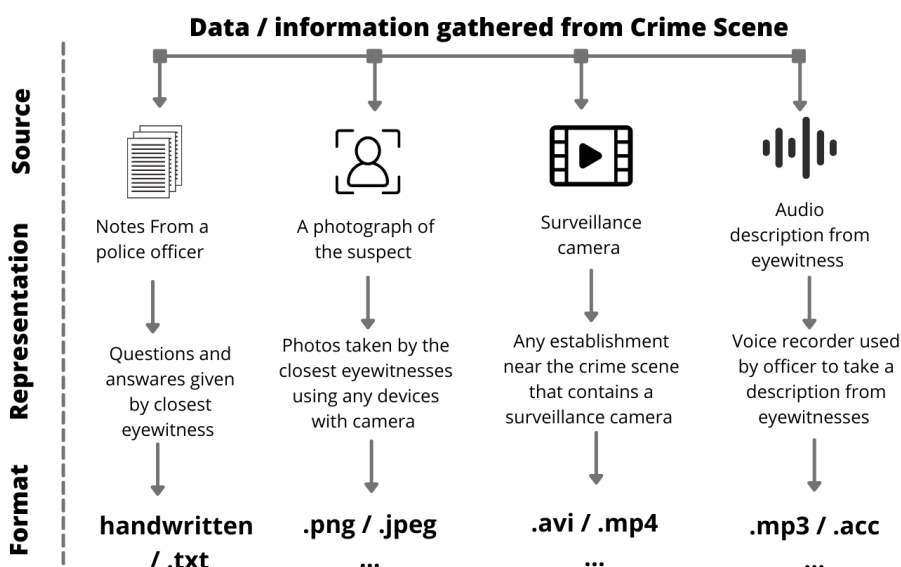


Figure 1. Type of information that can be collected from crime scene

overcome various challenges in text-to-image synthesis: Synthesizing high-quality images with a single generator-discriminator pair. As a result, for improved model stabilization, we used several loss functions, but the best results were produced using the hinge loss function and a one-stage training backbone based by unconditional image creation. This approach leverages deep text-image Fusion Blocks to combine sentence captions with graphical representations (DFBlock). Using Fusion Blocks, many affine Transformations and ReLU layers must be stacked to completely support text information. According to this approach, the gradient penalty takes matching into account and produces just one output direction that used to increase the generator's ability to synthesizes realistic material without the need for additional networks [20].

Lastly, Patashnik et al combine two existing models StyleGAN and CLIP for Text-Driven Manipulation of Imagery (StyleCLIP). This model is inspired from the capacity of StyleGAN to generate very realistic images; CLIP is used to adjust an input latent vector in response to a user's given text prompts in an optimization strategy. As a result, the model aims to provide a text-based interface for StyleGAN picture editing that does not necessitate such a large amount of time and effort to implement. In addition, it introduce the text-guided latent manipulation step, to be inferred for a given input picture using a latent mapper, which allows for quicker and more stable text-based manipulation [21].

So far, all text-to-image syntheses were based only on different GAN techniques, due to the absence of the following reasons in other approaches, no use of unsupervised approaches, no use of optimization technics during training, no model adaptation to new dataset and no generation of

new visual content [22]. However, Within the scope of this study, we are going to make use of a two-stage, a model base on an Quantized Variational Autoencoder in the first stage, and a transformer in the second [23]. Which will outperform GAN techniques, in terms of image quality and generation details.

C. Datasets used in our experiments

To be able to generate a complete facial image, we train several types of deep learning models, in addition, we use different datasets to serve both for training and testing. Table I. summarizes all these datasets, such as CelebA, CelebA-HQ, Flickr-Faces-HQ, Wider Face, and IAM Handwriting.

3. METHODOLOGY OF WORK

Our work concerns all types of data collected from the crime scene to generate a complete portrait of the suspect; Each type of data is fed and exploited into a specific deep learning network, as shown in Figure 2.

Source 1 - Written notes: we use the HTR deep network to transform handwritten characters to a digital text. We ameliorate the Handwritten Text Recognition (HTR) model, by adding some CNN layers trained on the IAM-HTR dataset [29], then we use it on the text description.

Source 2 – Picture of the suspect: Here, we opt for an attribute classification network. We choose to work with the ABM-CNN approach for multi-purposes, like high-level attributes classification and facial detection.

Source 3 – Video surveillance: In this case, we use a Yolov5 network to detect the suspect face in the video, we try to match it with our database, then we select a clean image to pass it to the attribute classification network.

TABLE I. Summary of the Datasets used in our work.

Ref	Dataset	Informations	The utility
[24]	CelebA	Over 200K celebrity photos with 40 attribute annotations make up this massive library of face characteristics.	Train attribute multi-label classification model, to detect them on face images.
[25]	CelebA-HQ	It contains over 30k images of a high-resolution face with descriptive text, for every image.	Face drawings and pictures can be generated with the use of text-to-image systems.
[26]	Flickr-Faces-HQ	This dataset contains 15k images high-quality PNG photos of 1024x1024 resolution, ranging in age, race, backdrop color, and text description.	Train the generator model.
[27]	Wider Face	Face detection benchmark dataset, with 32k images and labeled 393k	Train the Yolov5 model to detect and crop faces from video.
[28]	IAM Handwriting	It provides handwritten English text forms that may be used to train and test handwritten text recognizer and to make writer identification and verification tests.	Train a model capable of digitizing handwritten text descriptions of a potential suspect.

Source 4 – Speech Recognition Systems: Given the great evolution in Speech recognition systems from the leading companies in the market, we will try to use one of the existing models, justifying our choice based on accurate transcription services. To compare the three existing models like Sphinx-4, Microsoft Speech API and Google Speech API [30], G. Bohouta and V. Kępuska suggest utilizing a variety of audio recordings and compute the word error rate (WER) for each one of every service. The result of their experiment shows that the WER of the three aforementioned systems was acceptable; we find out that the Google API is the best one. For this reason, we use Google API as our Speech recognition model to transform audio information given by eyewitnesses into text description.

4. MODELS IMPLEMENTATION AND DISCUSSION

A. Hardware characteristics

For the test of our models, we use high-performance computing (HPC) infrastructure Cluster HPC-MARWAN:

- Compute Nodes: 2 * Intel Xeon Gold 6148(2.4GHz/20-core) / 192 GB RAM,
- GPU Node : 2 * NVIDIA P100 / 192 GB RAM,
- Storage Node: 2 * Intel Xeon Silver 4114(2.2GHz/20-core)/18 * SATA 6 TB.

B. Handwritten text recognition model

Regarding the HTR technology to read the handwriting from a notebook, the implementing technology uses 5 CNN

layers besides the input and output layers, the first layer is a filter kernel followed by a convolutional layer and a batch normalization layer, Finally, the non-linear RELU function and pooling layer is applied. The output of the CNN model has been directed towards 2 RNN (LSTM) layers, there are 256 features every timestep in the feature sequence, and the RNN uses this sequence to disseminate pertinent information. The RNN output sequence is mapped to a matrix of size 32x80. The last component is the Connectionist Temporal Classification (CTC) loss and decoding layer, The CTC calculates the loss value when training the NN with the RNN output matrix and the text that represents the ground truth. The CTC only receives the matrix during inference, and it decodes it to produce the final text. Each of these texts has a character limit of 32 characters (see Figure 3).

C. Text to face generation

Dataset preparation: We face in our work the problem of lacking facial datasets that regroups text description, real images, and sketched images. In order to resolve this issue, we train a model based on StyleGAN [31] to transform real images to sketches. We initialize two generators Gfrozen and pre-trained weights from a generator on images from the FFHQ source domain are used to train Gtrain. Gfrozen remains fixed throughout the process. However, the G trains are improved by optimization and an iterative layer-freezing approach. To retain the common latent space, the G train domain is moved according to a user-provided textual direction. We pass the multi-modal CelebA-HQ which contains over 30k images of the high-resolution face (see Figure 4).

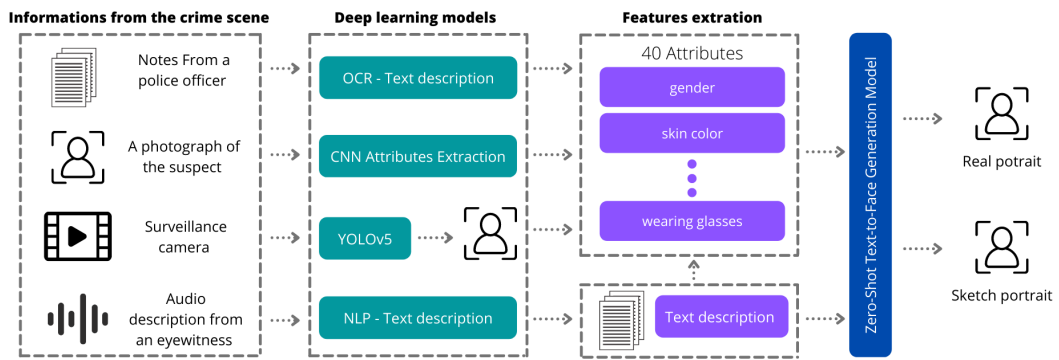


Figure 2. Illustration of our methodology to generate portraits

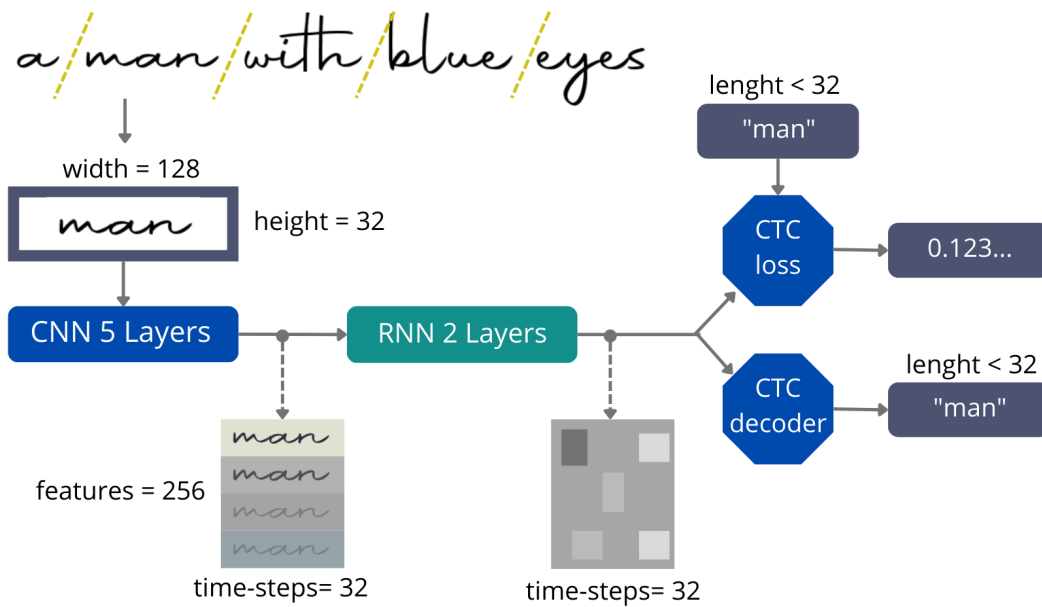


Figure 3. Model Overview of Handwritten Text Recognition (HTR)

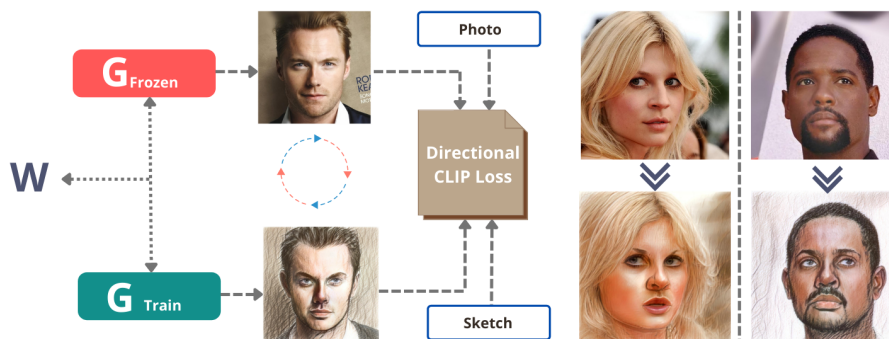


Figure 4. The setup we use to transform the real facial data to sketch

The architecture of the model: in the generation part, we choose the Zero-Shot Text-to-Image Generation based on an autoencoder who proved his superiority from other related work, in this area. We train it and adapt it to face images and sketch datasets. The model Zero-Shot Text-to-Image Generation consists of two-stage training Quantized Variational Autoencoder (VQ-VAE) and an Autoregressive transformer. The First Stage, it's for compressing each 256 x 256 RGB picture into a 32 by 32 grid of image tokens, each of which can have 8192 potential values, the model trains a Quantized Variational Autoencoder (VQ-VAE). This decreases the transformer's context size by 192 without sacrificing much visual quality as shown in Figure 5. In the second stage, the text and picture tokens are concatenated, and an autoregressive transformer is trained to represent the joint distribution of the two sets of data.

The Model input x is sent through an encoder to generate output; discrete latent variables $z_e(x)$ are then computed using a closest neighbor lookup Using the shared embedding space e as stated in Equation 1, the decoder's input is the equivalent embedding vector e_k as shown in Equation 2. The posterior categorical distribution $q(z|x)$ probabilities are defined as one-hot as follows:

$$q(z = k|x) = \begin{cases} 1 & \text{for } k = \text{armin}_j \|z_e(x) - e_j\|, \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

Where $z_e(x)$ is the output of the encoder network. We think of this model as a VAE in which $\log p(x)$ and ELBO can be constrained. A KL divergence constant equal to $\log K$ is obtained from the proposal distribution $q(z = k|x)$ by defining a simple uniform prior over z . Equation 1 and Equation 2 show how the representation is processed via the discretization bottleneck before being mapped onto the closest element of embedding e , as seen in the Figure 5.

$$z_q(x) = e_k, \text{ where } k = \text{argmin}_j \|z_e(x) - e_j\|, \quad (2)$$

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Handwritten text recognition model

Doubling the number of convolution layers allowed us to improve accuracy compared to the old model [29], it reaches an accuracy of 76% (2% more than the existing model. As shown in Figure 6, training a larger network is more efficient and quicker than training a smaller network. The model is training in 60 epochs, which stagnates around 55 epochs, also we can notice a slight improvement in loss. Our loss reaches 17%, while the loss from the state of art is 18%.

B. The YoloV5 Model

It is trained on WIDER FACE a face detection benchmark dataset that contains over 32k images label 393k faces. We split this dataset in our experiment as follows: 80% for training and 20% for testing, and perform running during 150 epochs. As shown in Figure 7, the training process

gives good results, concerning training and validation (the loss descends above 1.2%). In addition, the other metrics also performs well, we get for the Mean Average precision 93%, 90% for the Recall, and 88% for the Precision with an inference speed of 61 FPS.

C. Text-to-Face Model

To achieve the best performance, we test our model under multiple circumstances, changing the following parameters: dataset, data quantity, and data quality; each training took approximately 7 days trained under 200 epochs. Figure 8 presents four major model trainings; as we can see in Figure 8 - (a), the lack of quality and quantity, which leads to a blurry generation, does not respect the given caption reaching the highest loss of 63.4%.

In the second training Figure 8 - (b), we augment the quantity of data and keep the same quality of the image, we saw some improvements, but it still does not give the ideal generation, the loss reaches 56.2%. In the Figure 8 - (c), we switch to another dataset with high-quality images but low quantity data, the same problem arises here as in Figure 8 - (b). The final test in Figure 8 - (d) is done on CelebA-HQ with high quality and quantity images, we obtain promising results, both in real and sketch images reaching a loss of 37.61%. Through this testing, we learn that we must prioritize the good quality images over the quantity of data to achieve good results.

In our opinion, the model could be much accurate and achieve better results if we feed it with more high-quality images and descriptions. We put our two models of real images and sketch in a series of generation testing, to figure out how they react towards external descriptions of different people. In Figure 9, we choose different genders and skin colors; in this case, when it comes to accurately creating high-quality photos while maintaining the integrity of the text, both models perform admirably. Furthermore, we observe that the model generates accurate images when we consider all the 40 attributes provided by CelebA datasets.

6. PROPOSED IMPLEMENTATION

We propose a strategy to implement our model, the police officer collects text and audio from the eyewitness, using just a dashboard on a laptop or smartphone and an internet connection, the police officer could transform audio to text by using google API. Then prompt the text description of the suspect and send it to the datacenter. In another hand, the server receives the text description and passes it to the model to generate multi-faces and send it to the dashboard officer. On the spot the officer shows all virtual images to the eyewitness, the closest one will be sent to the server, to retrieve attributes and then re-send to the dashboard, the witness gives more detail to the police officer, about the offender, and the police officer sends the prompt text to the server, and so one of the manipulations is repeated until the witness valid the portrait of the suspect, see Figure 10.

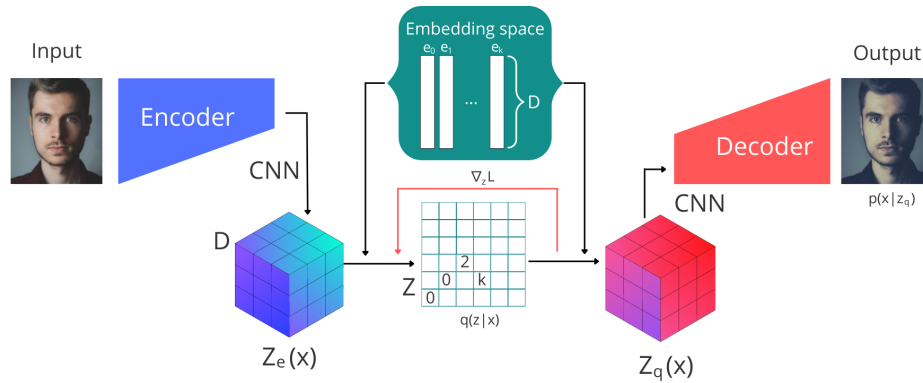


Figure 5. Model Overview of Quantized Variational Autoencoder (VQ-VAE)

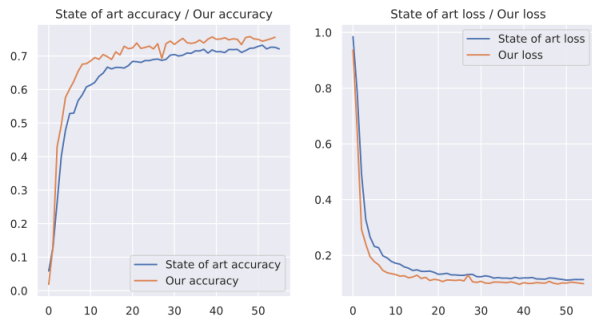


Figure 6. Results of handwriting model training for both state of art and ours result based on accuracy and loss

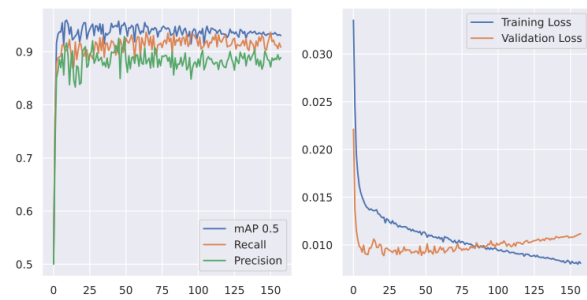


Figure 7. Results of YoloV5 face detection according to different metrics



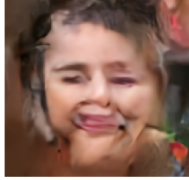


<p>(a)</p> <p>Dataset : CelebA</p> <p>Number of images and text for training: 10k</p> <p>Training loss reach: 63.4%</p> <p>Sample :</p> <p>The woman opening her mouth slightly has blond hair.</p> 	<p>(b)</p> <p>Dataset : CelebA</p> <p>Number of images and text for training: 200k</p> <p>Training loss reach: 56.2%</p> <p>Sample :</p> <p>the woman opening her mouth slightly has earrings .</p> 	<p>(c)</p> <p>Dataset : FFHQ</p> <p>Number of images and text for training: 15k</p> <p>Training loss reach: 50.15%</p> <p>Sample :</p> <p>This girl is about 15 to 20 years old and has big bright brown eyes with double eyelids .</p> 	<p>(d)</p> <p>Dataset : CelebA - HQ / REAL and SKETCH</p> <p>Number of images and text for training: 30K</p> <p>Training loss reach: 37.61%</p> <p>Real images sample :</p> <p>this is a man with a high nose and a pair of big blue eyes , and his nose is small .</p>  <p>Sketch images sample :</p> <p>The woman with black eyebrows and big eyes has a pair of brown eyes .</p> 
---	---	---	---

Figure 8. Training sample of the generator model under different circumstances

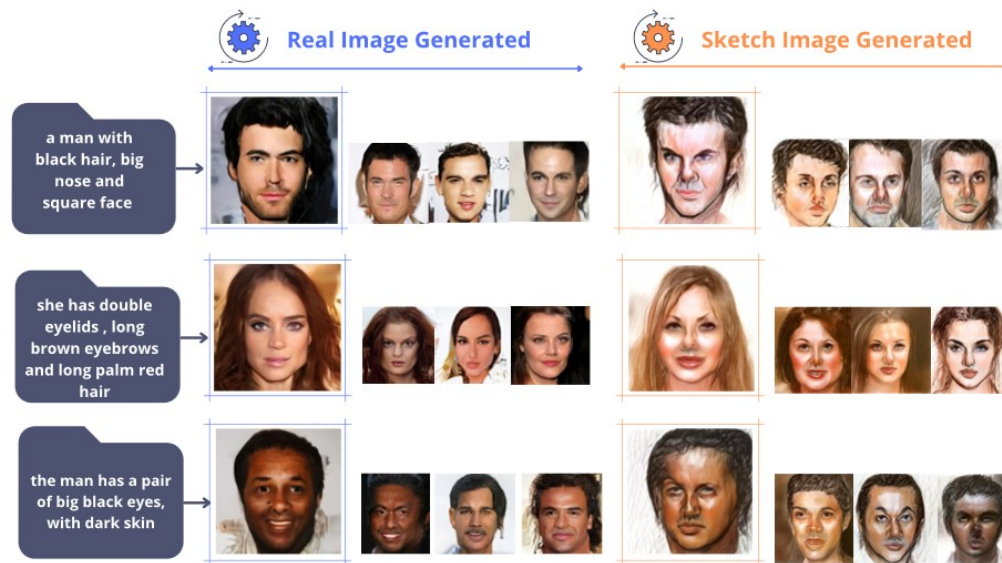


Figure 9. Image generation based on text description from both real and sketch models

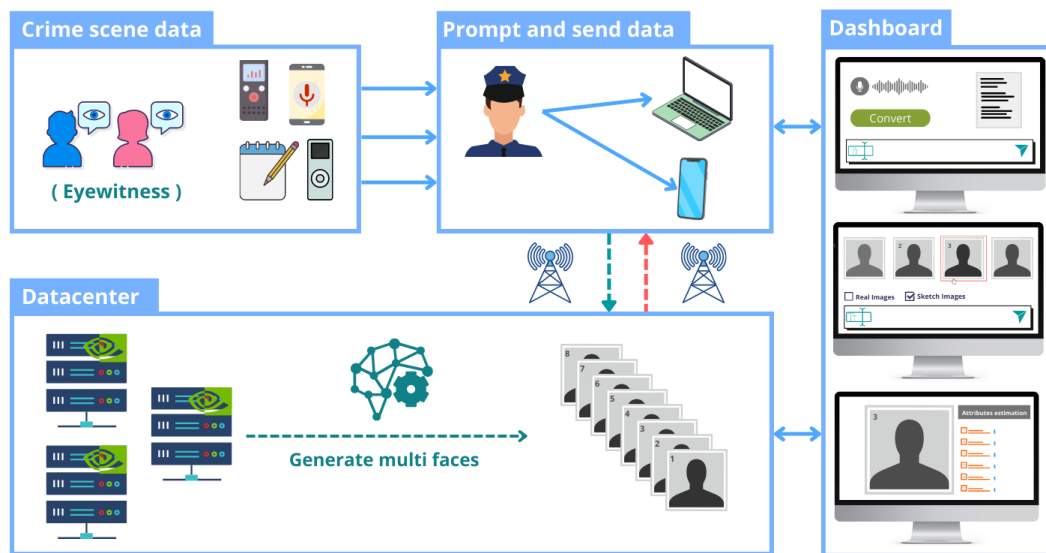


Figure 10. Proposed implementation

7. CONCLUSION

This work represents an overview of deep learning models that can be of a great usefulness for law enforcement to generate two types of facial portraits: real and sketch. To this end, we collect four types of data from the crime scene: handwritten text and audio from officer notes taken according to an eyewitness, images from smartphones, and video from the phones or surveillance cameras. Then, we use two pre-trained models; the first one, ABM-CNN, is for making attribute multi-label classification in order to get all estimated attributes on given images; and the second model, Google Speech API, is used to transform speech to text.

We also train three models: the first one, the handwritten model, is trained on the IAM handwriting dataset, targeting so to read notes and digitalizing the text; it reaches an accuracy of 76%, outperforming state-of-the-art results. For the second one, we train YoloV5 using Wider Face dataset, to detect one or multiple faces within images or videos, reaching a Mean Average precision of 93%, a Recall of 90%, and a Precision of 88%. In the third model, we adapt Zero-Shot-Text-to-Image to generate faces by training it over two stages; the first stage aims to train a Quantized Variational Autoencoder (VQ-VAE), then the second stage looks up to train an Autoregressive transformer.



The obtained results with this model outperform those of the existing models, and allows us generating accurate facial images with a loss of 37.4%. The combination of all these models together leads us to build a complete system that can collect data, reformat it, find faces, estimate attributes, then generate portraits.

REFERENCES

- [1] M. Boukabous and M. Azizi, "Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, pp. 1131–1139, feb 2022. [Online]. Available: <http://ijeeecs.iaescore.com/index.php/IJEECS/article/view/26734>
- [2] C. D. Frowd, M. Pitchford, F. Skelton, A. Petkovic, C. Prosser, and B. Coates, "Catching even more offenders with EvoFIT facial composites," in *Proceedings - 3rd International Conference on Emerging Security Technologies, EST 2012*, 2012, pp. 20–26.
- [3] I. Idrissi, M. Azizi, and O. Moussaoui, "A Lightweight Optimized Deep Learning-based Host-Intrusion Detection System Deployed on the Edge for IoT," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 209–216, 2022.
- [4] M. Allison, C. Basquin, J. G. A. P. in Criminal, and undefined 2017, "Assessing the Accuracy of English-As-a-Second-Language Eyewitness Testimonies and Contemporaneous Officer Notes Using Two Methods," *dev.cjcenter.org*, no. 1, p. 13, 2017. [Online]. Available: http://dev.cjcenter.org/_files/apcj/APCJSRING2017-allison.pdf_1495139764.pdf
- [5] Y. Lin, K. Fu, S. Ling, J. W. I. T. on ..., and undefined 2021, "Toward Identity Preserving Face Synthesis Between Sketches and Photos Using Deep Feature Injection," *ieeexplore.ieee.org*. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9411729>
- [6] M. Berrahal and M. Azizi, "Improvement of facial attributes' estimation using Transfer Learning." Institute of Electrical and Electronics Engineers (IEEE), mar 2022, pp. 1–7.
- [7] M. Berrahal and M. Azizi, "Optimal text-to-image synthesis model for generating portrait images using generative adversarial network techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 2, pp. 972–979, feb 2022. [Online]. Available: <http://ijeeecs.iaescore.com/index.php/IJEECS/article/view/26824>
- [8] M. Berrahal and M. Azizi, "Augmented binary multi-labeled CNN for practical facial attribute classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 973–979, aug 2021. [Online]. Available: <http://ijeeecs.iaescore.com/index.php/IJEECS/article/view/24782>
- [9] M. Boukabous and M. Azizi, "Multimodal Sentiment Analysis using Audio and Text for Crime Detection." Institute of Electrical and Electronics Engineers (IEEE), mar 2022, pp. 1–5.
- [10] T. Karras NVIDIA and S. Laine NVIDIA, "A Style-Based Generator Architecture for Generative Adversarial Networks Timo Aila NVIDIA." Tech. Rep. [Online]. Available: <https://github.com/NVlabs/stylegan>
- [11] T. C. Wei, U. Sheikh, and A. A. H. A. Rahman, "Improved optical character recognition with deep neural network," *Proceedings - 2018 IEEE 14th International Colloquium on Signal Processing and its Application, CSPA 2018*, pp. 245–249, may 2018.
- [12] B. Balci, D. Saadati, D. S. f. V. R. ..., and undefined 2017, "Handwritten text recognition using deep learning," *cs231n.stanford.edu*. [Online]. Available: <http://cs231n.stanford.edu/reports/2017/pdfs/810.pdf>
- [13] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer, "Facial Attributes Classification using Multi-Task Representation Learning," Tech. Rep.
- [14] A. Kherraki and R. El Ouazzani, "Deep convolutional neural networks architecture for an efficient emergency vehicle classification in real-time traffic monitoring," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 110–120, mar 2022. [Online]. Available: <https://ijai.iaescore.com/index.php/IJAI/article/view/21104>
- [15] C. Park, J. Seo, S. Lee, C. Lee, H. Moon, S. Eo, and H.-S. Lim, "BTS: Back TranScription for Speech-to-Text Post-Processor using Text-to-Speech-to-Text," pp. 106–116, jul 2021. [Online]. Available: <https://aclanthology.org/2021.wat-1.10>
- [16] I. J. 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*, 2014.
- [17] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," vol. 2017-October, pp. 5908–5916, dec 2016. [Online]. Available: <https://arxiv.org/abs/1612.03242v2>
- [18] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1947–1962, oct 2017. [Online]. Available: <http://arxiv.org/abs/1710.10916>
- [19] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," Tech. Rep. [Online]. Available: <https://github.com/taoxugit/AttnGAN>.
- [20] M. Tao, H. Tang, S. Wu, N. Sebe, X.-Y. Jing, F. Wu, and B. Bao, "DF-GAN: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis," *IEEE Transactions on Multimedia*, aug 2020. [Online]. Available: <http://arxiv.org/abs/2008.05865>
- [21] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery," pp. 2065–2074, mar 2021. [Online]. Available: <https://arxiv.org/abs/2103.17249v1>
- [22] M. Berrahal and M. Azizi, "Review of DL-Based Generation Techniques of Augmented Images using Portraits Specification," in *4th International Conference on Intelligent Computing in Data Sciences, ICDS 2020*. Institute of Electrical and Electronics Engineers (IEEE), nov 2020, pp. 1–8.
- [23] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation," feb 2021. [Online]. Available: <https://arxiv.org/abs/2102.12092v2>
- [24] "Large-scale CelebFaces Attributes (CelebA) Dataset." [Online]. Available: <http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>
- [25] "CelebAMask-HQ Dataset." [Online]. Available: <https://mmlab.ie>

cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html

- [26] “NVlabs/fhq-dataset: Flickr-Faces-HQ Dataset (FFHQ).” [Online]. Available: <https://github.com/NVLabs/fhq-dataset>
- [27] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A Face Detection Benchmark,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 5525–5533, nov 2015. [Online]. Available: <https://arxiv.org/abs/1511.06523v1>
- [28] “Build a Handwritten Text Recognition System using TensorFlow — by Harald Scheidl — Towards Data Science.” [Online]. Available: <https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow>
- [29] “Build a Handwritten Text Recognition System using TensorFlow — by Harald Scheidl — Towards Data Science.” [Online]. Available: <https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow>
- [30] V. Képuska, “Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx),” *International Journal of Engineering Research and Applications*, vol. 07, no. 03, pp. 20–24, mar 2017.
- [31] R. Gal, O. Patashnik, H. Maron, G. Chechik, and D. Cohen-Or, “StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators,” aug 2021. [Online]. Available: <https://arxiv.org/abs/2108.00946v2>



Mohammed BERRAHAL is a Ph.D. candidate in computer engineering at Mohammed First University in Oujda, Morocco, where he does research on Deep Learning security and law enforcement applications. He earned a Master of Science in internet of things from ENSIAS, Mohammed 5 University in Rabat, Morocco (2018) and a Bachelor of Science in computer engineering from ESTO, Mohammed First University in Oujda, Morocco (2017). He is further qualified in artificial intelligence, 3D modeling, and programming. In addition, he has reviewed for a number of international conferences and publications. And is employed as an administrative assistant at Mohammed First University at the present time. He can be reached by email at m.berrahal@ump.ac.ma.



Prof. Dr. Mostafa AZIZI received a State Engineer degree in Automation and Industrial Computing from the Engineering School EMI of Rabat, Morocco in 1993, then a Master degree in Automation and Industrial Computing from the Faculty of Sciences of Oujda, Morocco in 1995, and a Ph.D. degree in Computer Science from the University of Montreal, Canada in 2001. He earned also tens of online certifications in Programming, Networking, AI, Computer Security ... He is currently a Professor at the ESTO, University Mohammed 1st of Oujda. His research interests include Security and Networking, AI, Software Engineering, IoT, and Embedded Systems. His research findings with his team are published in over 100 peer-reviewed communications and papers. He also served as PC member and reviewer in several international conferences and journals. He can be contacted at email: azizi.mos@ump.ac.ma.