Sketch guided face synthesis using conditional variational autoencoder

Edson Odake and Eduardo Parente Ribeiro

***Abstract*— Forensic sketches, often the only visual leads in criminal cases, typically lack the detail and realism that can help in public identification tasks. This paper presents a Conditional Variational Autoencoder (CVAE) approach for transforming forensic sketches into photorealistic facial images. Using a stochastic edge map extraction of images from a dataset, our model bypasses the need for manually paired sketch-photo databases, enhancing scalability. The model was evaluated on several metrics, demonstrating the capability of working on different image styles. When using the Facenet, the similarity of generated images using the CVAE is 56.8% better than the simpler AE.**

***Keywords*— Conditional Variational Autoencoder, Photorealistic Face Synthesis, Forensic Sketch, Image-to-Image Translation**

I. Introduction

Forensic sketches, traditionally created from eyewitness descriptions, are vital when direct information about a suspect, such as fingerprints or names, is unavailable. These sketches are typically compared against a database either manually by experts or via face recognition technology. However, converting these sketches into photorealistic images can enhance accuracy and public engagement, leading to increased community tips and suspect identification. Photorealistic images integrate more seamlessly with digital technologies and databases, including facial recognition software.

Despite advances in generative algorithms for creating images from facial sketches, current methods often rely on limited databases pairing sketches with original photos [1], [2], a practice that lacks scalability and restricts style diversity. Furthermore, existing models trained on unpaired images are primarily evaluated on artificial or digital sketches [3], [4]. This evaluation fails to account for non-digital, hand-drawn sketches, which limits the accuracy of the generated images across more traditional forms of sketching. Another significant gap in the literature is the absence of specific metrics to assess the fidelity of the generated facial features to their original counterparts.

This paper presents a generative model that transforms facial sketches into photorealistic images. Unlike traditional approaches, our model uses a stochastic edge map extraction, avoiding the constraints of a narrow style range and enhancing generalizability. We also provide a novel architecture combining a Conditional Variational Autoencoder (CVAE) with skip connections [5] and attention mechanisms [6] to improve the generated image quality. Utilizing a CVAE, the generated image captures essential conditional information such as skin and hair color, compensating for details not provided by the sketch.

Our model is trained on CelebA [7] database and validated against a variety of previously unseen hand-free sketches, ensuring robust performance across different artistic styles representation. We evaluate the images using several perceptual metrics to verify the preservation of facial features. This methodology broadens the practical applicability in fields such as forensic and artistic uses, where flexibility in handling different sketch styles is crucial.

II. Related work

The synthesis of photos guided by sketches is essentially an Image-to-Image problem, which tries to optimize the mapping between two domains. The most commonly used models for this task are the VAE and GANs [8].

Nastaran Moradzadeh Farid et al [9] developed a sketch to image GAN with impressive results, but the evaluation was only applied to a single-style database. Yongyi Lu et al [10] proposed a Contextual GAN using the sketch as a weak contextual constraint. This method proved robust when applied to ugly sketches but did not guarantee the identity preservation of the image. Phillip Isola et al [11] provided a Conditional GAN using a U-Net-based generator to translate multiple styles of images, however, they focused on more generic images.

Several GAN-based models were proposed more specifically for converting sketches to facial photos [12], [13], [14], but their evaluation were limited to the same database used to train the model. Jun Yu et al [15] went one step further and evaluated the model on a variety of sketch styles, outside the training database, which provided blurry but interesting results. However, all of these models require paired images for the training process. Mingming Hu [13] used the xDog filter [16] to generate sketches from a facial photo and used it as training input, which provided good results but didn't evaluate the model on real sketches. Other authors [17], [18] employed the powerful StyleGAN pre-trained model and created a latent space mapping using as input a sketch and a text description ensuring appealing results, however, the evaluation was also limited to the same database used for training.

The different styles between the training database and sketches produced by several artists present a challenge for generating a good image based on sketches. Yuhang Li [3] achieved good results on ugly digital sketches. He approached the problem using different methods to synthesize sketches from the original photo. Shu-Yu Chen [4] provided an architecture centered on generating the facial components separated (eyes, mouth, nose), by applying windows to encode and decode these features, achieving impressive results on free-hand

sketches. However, their evaluation was limited to digitally drawn sketches.

Xing Di [19] used a complex architecture combining CVAE and GAN in a three-stage training process to achieve facial photos from sketches. However, GAN-based models are difficult to train because of convergence problems and mode collapse. In this work, we propose a simpler architecture using a more robust CVAE with a more diverse preprocessing step, making the model more suitable to achieve good results on different style images.

III. Methodology

The proposed CVAE (Fig. 1) receives a 64x64 edge map as input along with an array of attributes describing the face's features. The model uses skip connections [5] to preserve both semantic and structural information about the initial image. Residual connections are also implemented to improve the model training [20].

An attention cell is added to the encoder before concatenating the skip connections with the encoder, to ensure that the semantic and structural information are well aligned [6]. The conditional information is positioned in both, the encoder and decoder network, mostly to give information such as hair and skin color.

The objective function is inspired by Irina Higgins work [21], which provides a simple way to adjust the latent space entanglement. The reconstruction loss chosen is the structural similarity index measure (SSIM) [22] which provided better empirical results when compared with pixelwise metrics like MSE and PSNR.

To obtain a diverse training database a preprocessing pipeline was created, producing different edge maps from the original images. The parameters of the pipeline are set stochastically to guarantee data diversity and the results are evaluated across different unseen databases.

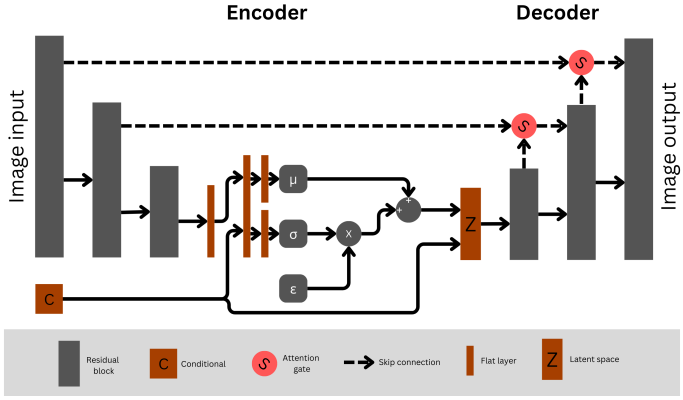


Fig. 1: Final model architecture.

The main components of the proposed CVAE are:

Residual block: Technique used to prevent vanishing gradient by summing the input with the convolution output, improving deep model training.

Conditinal input: Contains 14 attributes selected from CelebA database to control high-level features, like hair, skin color, and smile.

Skip connections: This method is used to concatenate the encoder's feature maps with the decoder's, providing structural information while generating the output image.

Attention gate: Regularization method that combines the skip connection with the decoder's previous layer to reconstruct the next layer.

Flat layer: The flatten operation is applied after the convolutional step to get a one-dimensional array.

Latent space: Contains low dimensional features used to reconstruct the output image.

A. Preprocessing pipeline

The first step is to extract edges from the images using the OpenCV canny detector [23]. The extracted edges are often not continuous and present several gaps. A morphological closing operation (dilatation and erosion) is applied on the edge map to fill these gaps while keeping the shape and width.

At last, the image passes through a Gaussian filter to thicken the edges providing a more diverse edge width. Then, the pixels are converted to a binary representation. The binarization is useful to standardize stroke styles on the sketch. The parameters of each step of the preprocess are selected stochastically by randomly defining the thresholds of the canny detector and the size of the Gaussian filter kernel for each image, providing a diverse representation of the edge map. Since the CVAE input is 64x64, the image is resized to fit this shape.

B. CVAE Objective Function

An autoencoder (AE) consists of an encoder followed by a decoder. The encoder maps high-dimensional data into a low-dimensional space, maintaining important features. The decoder takes this low-dimensional data and learns the reverse map to the high-dimensional space.

The objective of the autoencoder is to achieve a perfect reconstruction as output. There is no control over the latent space, resulting in a non-continuous, entangled, and messy space. The variational autoencoder (VAE) was introduced to add regularization in the latent space, providing a smoother and more stable space.

The Conditional VAE (CVAE) provides a secondary input type, which is concatenated with the encoder and decoder. If you're working with animal image generation, the conditional may be a one-hot encoded label representation, containing the animal species. Since the sketch doesn't provide color information, it's passed through a conditional input.

Figure 1 shows a CVAE abstraction. The CVAE loss function is given by [17]:

$$L = \frac{1}{2}\sum_{i} 1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2 - \mathbb{E}_{z\sim q(z|x_i)}[\log p(x_i|z)] \tag{1}$$

In this equation, L is the total loss function, μ_i and σ_i are the mean and standard deviation of the Gaussian distribution representing the encoder output. The posterior distribution $q(z|x_i)$ represents the encoder mapping from $\mathbf{X}$ domain to

the $\mathbf{Z}$ latent space, given a sample x_i. The decoder likelihood distribution approximation is represented by $p\ (x_i|z)$.

The first component is the regularization term. It measures the Kullback-Leibler (KL) divergence between the encoder's posterior distribution $q\ (z|x_i)$ and the prior distribution $p(z)$, which is a standard normal distribution (= 0, = 1). The second component is the reconstruction term. It represents the expected log-likelihood of reconstructing the input x_i from the latent variable z using the decoder. Here, $\mathbb{E}_{z\ q\ (z|x_i)}$ denotes the expected value with respect to the distribution $q\ (z|x_i)$.

The regularization component is multiplied by to balance the trade-off between the latent space constraint and the reconstruction accuracy in the CVAE [21]. This allows us to control the emphasis placed on the regularization term versus the reconstruction term. Additionally, the reconstruction component is replaced by the SSIM loss as its closed form [22]. Since the AEs only present the reconstruction loss, the is not applied, but the same SSIM loss is used as the reconstruction component.

$$\text{SSIM}(x;y) = \frac{(2\ _x\ _y + c_1)(2\ _{xy} + c_2)}{(\ _x^2 + \ _y^2 + c_1)(\ _x^2 + \ _y^2 + c_2)}; \qquad (2)$$

Where c_1 and c_2 are small constants that stabilize the division with weak denominators.

C. Evaluation Metric Overview

A commonly used metric to evaluate how close two images are is the mean square error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i \quad y_i)^2 \qquad (3)$$

The SSIM can also be used. It is a perception-based model that incorporates luminance and contrast masking. These methods are effective when the images are properly aligned. Images generated using the edge detection from image dataset are well evaluated with these simple metrics since the target image is perfectly aligned with the edge map. However, images generated with hand sketches are more difficult to evaluate with these metrics. MSE is sensible to the slight change in pixel position and the SSIM may get disturbed by color differences, like the background.

The FID metric is chosen to provide an evaluation more focused on image features instead of pixel-wise details. This metric uses Inception V3 to extract a feature vector from the image and calculate the Frechét Distance. FID is widely used for evaluating the quality of images generated by GANs.

A more specific metric can be defined using Google's FaceNet model developed for face recognition. This model provides state-of-the-art results in image recognition, being a good choice for evaluating the similarity of the generated and original faces.

D. Databases

The model is trained using the CelebA dataset [7], containing 202.599 facial images each image having 40 attributes. The model is trained using 50k images and only attributes containing color information are used. The images have a 218x178 dimension. Since the CelebA dataset doesn't contain hand-drawn sketches it would be unfair to evaluate the model only using this base.

Two different styles of images are used to evaluate the model and ensure good performance under unseen samples:

CelebA: This dataset is used to check the performance under unseen images with a distribution similar to the training data;

CUHK Student database: Includes the sketch of 188 asian students paired with the original photo. The sketch dimension is 200x250;

IV. Results

We provide a comparison between the proposed models and a simpler implementation to justify the complexity. Figure 2 illustrates the training and validation loss across different models. It's important to note that the CVAE loss is different from the AE since it also presents a regularization component.

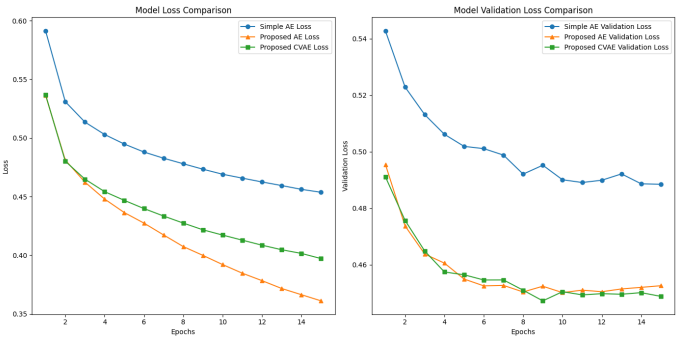


Fig. 2: Training and validation history.

Figure 3 provides the quantitative results, enforcing the superior result of the proposed over the simpler model. Figure 4 shows the original image, the extracted edges, and the image generated using different models. The Simpler AE model provides blurry and generic images which is improved by using the Proposed AE and CVAE.

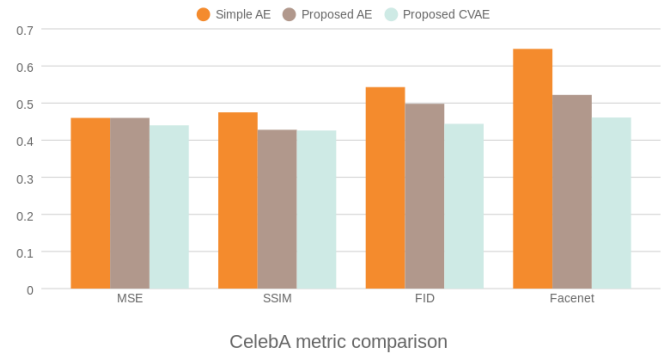
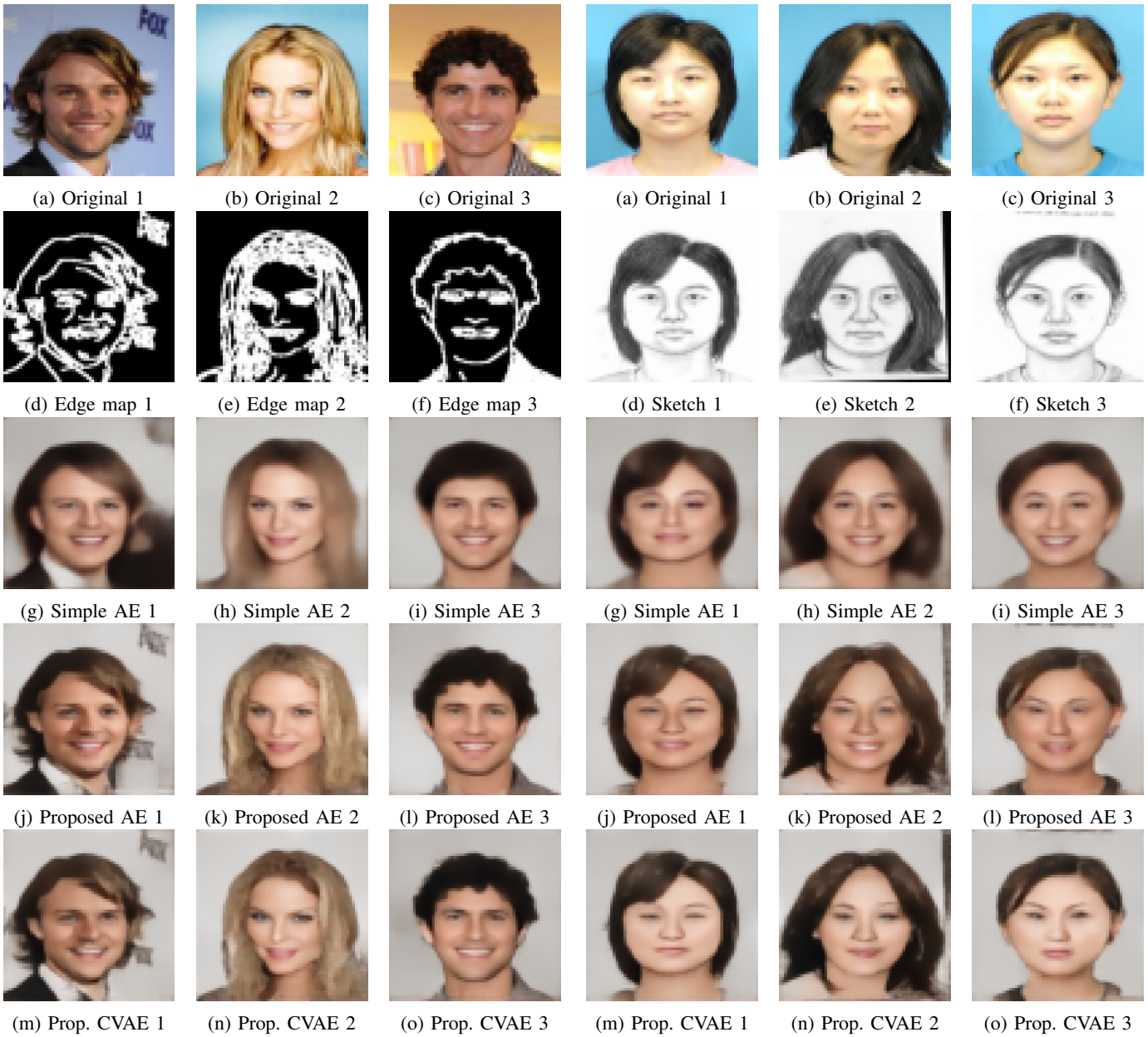


Fig. 3: CelebA performance metric comparison for each model on 10 samples (Metrics were scaled to the same perspective).

The same procedure is applied to the CUHK Student Base. Figure 5 shows the importance of having conditional input to



(a) Original 1 (b) Original 2 (c) Original 3
(d) Edge map 1 (e) Edge map 2 (f) Edge map 3
(g) Simple AE 1 (h) Simple AE 2 (i) Simple AE 3
(j) Proposed AE 1 (k) Proposed AE 2 (l) Proposed AE 3
(m) Prop. CVAE 1 (n) Prop. CVAE 2 (o) Prop. CVAE 3

Fig. 4: CelebA image comparison

(a) Original 1 (b) Original 2 (c) Original 3
(d) Sketch 1 (e) Sketch 2 (f) Sketch 3
(g) Simple AE 1 (h) Simple AE 2 (i) Simple AE 3
(j) Proposed AE 1 (k) Proposed AE 2 (l) Proposed AE 3
(m) Prop. CVAE 1 (n) Prop. CVAE 2 (o) Prop. CVAE 3

Fig. 5: CUHK image comparison

generate a more accurate color reconstruction. MSE, SSIM, and FID provided similar results but the Facenet comparison demonstrated the superiority of the proposed models over the Simpler AE (Figure 6).

Figure 7 shows the transition while interpolating hair and skin color attributes to illustrate the impact of the conditional input on the generated images. The model demonstrates the ability to change these features without altering other aspects of the image.

The models were structured to have similar depths and parameters, ensuring a fair comparison. Each model consisted of multiple convolutional layers with 3x3 kernel sizes, followed by batch normalization and ReLU activation functions to introduce non-linearity. The simpler AE architecture included 13 layers. While sharing a similar core structure, the proposed AE and the CVAE architectures included additional layers for residual connections, attention mechanisms, and conditional input processing, bringing the total to 26 layers. These additions primarily serve to improve regularization. All models were trained for 15 epochs (without early stopping) with a batch size of 64, using the ADAM optimizer with a learning rate of 0.001. Batch normalization was applied after each convolutional layer to mitigate issues like exploding gradients.

V. Conclusions

We developed a CVAE to generate images based on facial sketches and a few attributes. The model was trained only on artificially synthesized sketches using edge detection techniques among other preprocessing methods, and evaluated across different style sketches. Despite being trained on