A Two-Stage Emotion Generation Model Combining CGAN and pix2pix

Yuanqing Wang, Zhoukou Normal University, China & City University, Malaysia*
Dahlan Abdul Ghani, Malaysian Institute of Information Technology, University Kuala Lumpur, China
Bingqian Zhou, Zhoukou Normal University, China & Universiti Malaysia Sarawak, Malaysia

ABSTRACT

Computer vision has made significant advancements in emotional design. Designers can now utilize computer vision to create emotionally captivating designs that deeply resonate with people. This article aims at enhancing emotional design selection by separating appearance and color. A two-stage emotional design method is proposed, which yields significantly better results compared to classical single-stage methods. In the Radboud face dataset (RaFD), facial expressions primarily rely on appearance, while color plays a relatively smaller role. Therefore, the two-stage model presented in this article can focus on shape design. By utilizing the SSIM image quality evaluation index, our model demonstrates a 31.63% improvement in generation performance compared to the CGAN model. Additionally, the PSNR image quality evaluation index shows a 10.78% enhancement in generation performance. The proposed model achieves superior design results and introduces various design elements. This article exhibits certain improvements in design effectiveness and scalability compared to conventional models.

KEYWORDS
CGAN, GAN, Image to Image Translation, Visual Communication Design

1. INTRODUCTION

AI-Generated Content (AIGC) (Liu, 2023; Kang, 2023) has made significant progress in content generation. Designers can now use AIGC to create emotionally captivating designs that resonate deeply with people. This paper utilizes deep learning technology to enhance the selection of emotional design through the separation of appearance and color. Deep learning technology can solve many problems, such as object detection, image classification, image segmentation (Ranjbarzadeh & Tirkolaee, 2023), and image generation, and has been widely applied in various industries (Weber, Arabnia, Aydin, Tirkolaee, 2023; Song, Li, et al., 2022; Wu, 2022; Morales & Suárez-Rocha, 2022; Dayyala, 2022). Deep learning (Miao & Ruomu, 2023; Zhou & Fang, 2022) is a machine learning subfield originating from the study of artificial neural networks. Currently, common neural networks used in deep learning include CNN, RNN, DNN, PNN, and others. These algorithms effectively model...
complex relationships within data through multi-layer representations. Unsupervised learning (Kumar, et al., 2021) has always posed a significant challenge for researchers. However, the rapid progress in unsupervised learning can be attributed to extensive research on generative models. These models can generate new samples based on the high-dimensional data distribution. In 2014, Ian Goodfellow proposed Generative Adversarial Networks (GANs) (Dalva, et al., 2022; Yang, et al., 2022; Kang, et al., 2023; Goodfellow, 2016; Li, et al., 2021; Zou & XiuFang, et al., 2022) which outperformed other generative models, producing better samples owing to the generation of countermeasures network. Compared to Pixel RNN (GANs) (van den Oord, et al., 2016; Salimans & Kingma, 2016), GAN generates a sample faster. In contrast to Variational Auto-Encode (VAE) (Sahu & Majumdar, 2021; Zhu, et al., 2021), a GAN does not have a lower bound of change, and its various adversarial generation networks gradually converge, while VAE exhibits some bias. Furthermore, a GAN does not possess a lower variation bound or a complex partition function like deep Boltzmann machines (Liu, et al., 2021; Luo, et al., 2021). GANs can generate samples simultaneously, eliminating the need for repeatedly applying the Markov chain solver (Karimi, et al., 2021; López-Suárez & Urriza, 2021). By comparison, it was found that GAN is more suitable for the application of this paper. Hence, this paper mainly focuses on generating content design, which requires diversification.

During the training process of a GAN, the generator network learns the distribution of real samples and determines whether an input image or real data is generated. The optimization of the entire model involves solving a “binary minimax game” problem. Currently, GANs represent the state-of-the-art in generative models. As researchers delve deeper into the realm of GANs, they have proposed several improved algorithms and innovative applications. For instance, the WGAN (Arjovsky, et al., 2017; Gulrajani, et al., 2017; Li, et al., 2021; Dong, et al., 2020; Xu, et al., 2021; Wang, et al., 2021) addresses the instability issues encountered during GAN training. DCGAN (Radford, et al., 2015; Deleforge, et al., 2021; Kim, et al., 2021) utilizes CNN in GAN to produce images of better quality. EBGAN (Chen, et al., 2016) treats the discriminator as an energy function, assigning low energy to actual data and high energy to fake data. On the other hand, InfoGAN (Liu & Lu, 2021; Gao, et al., 2021; Chen, et al., 2021) achieves decomposable feature representation through unsupervised learning. It combines a GAN with maximizing the mutual information between the generated image and the input code.

The Conditional Generative Adversarial Network (CGAN) (Kwon, et al., 2022; Xie, et al., 2022; Jabbar, et al., 2023; Ruan, et al., 2023; Zhang, et al., 2022) and CycleGAN (Wang, et al., 2023; Chen, et al., 2022; Zhang, et al., 2022) are powerful tools for emotional design, enabling designers to harness the emotional impact of images and create designs that deeply resonate with people. Whether it involves crafting personalized avatars or transforming real-world environments into virtual realms, CGAN and CycleGAN empower designers to produce emotionally compelling designs that resonate with individuals’ hearts and minds. As designers continue to explore the possibilities offered by these technologies, we anticipate a new era of emotionally engaging design that transforms our relationship with technology and each other. CGAN and CycleGAN find broad applicability in image design, spanning diverse domains such as fashion design, interior design, and video game development.

CycleGAN and pix2pix (Min & Kim, 2022; Ding, et al., 2022; Zhang, et al., 2023; Shen, et al., 2023) are two distinct types of Generative Adversarial Networks (GANs) employed for image-to-image translation tasks. While they share the common goal of image translation, their approach and specific applications differ.

However, the aforementioned networks primarily function as end-to-end design methods, typically addressing a single aspect, and may lack the necessary design elements for effective emotional design (Ng, et al., 2021). Consequently, this paper proposes a two-stage emotional design model to overcome the limitations of a single model. By combining Conditional Generative Adversarial Nets and pix2pix, a two-stage emotion generative model is developed based on the characteristics of the emotional design field. In the domain of visual communication, contour, and color are the primary visual factors that influence emotions. Contour refers to the lines forming the outer edges of shapes
or objects, while color encompasses the attributes that enrich the content of the lines. This article uses conditional generation to produce contour attributes, allowing adversarial networks to generate images while focusing solely on contour information. Subsequently, the pix2pix network is used to enhance the color attributes of the images. This comprehensive design leverages computer vision technology to divide emotion generation into two steps. First, the CGAN generates contour information enabling the network to concentrate exclusively on contour generation without being influenced by color information. Second, the pix2pix network is employed to generate color information, emphasizing the acquisition of color attributes since contour information remains unchanged. Experimental results demonstrate that using a two-stage emotional design approach yields superior images compared to generating them in a single step.

To sum up, the proposed method introduces the following key innovations:

1. An image contour extraction algorithm is used to generate a training set and perform dimensionality reduction on the image, resulting in the extraction of image contours.
2. An image translation network that enriches contour images with color information.

The rest of this paper is organized as follows. First, relevant literature is introduced in Section II. Then, the proposed two-stage emotion generation model is presented in Section III. The experimental results are reported in Section IV, and conclusions are drawn in Section V.

2. RELATED WORKS

The GAN comprises two models: the generative and discriminant models. The generative model catches the distribution of sample data, while the discriminative model is a binary classifier. The input to the generative model is a random noise that follows a simple distribution, and the output is a generative image with the same size as the training image. These generated samples are then fed to the discriminative model. The generative model produces outputs with low probabilities (indicating that they are generated samples), while the goal of the generative model is to deceive the discriminative model into outputting high probabilities (mistakenly identifying them as real samples), leading to a competitive and adversarial relationship. Compared to other generative models, GANs have been proven to be the most effective, surpassing models like VAEs. Although VAEs can generate high-quality samples, their produced outputs often lack realism compared to GANs. VAEs are better suited for tasks such as data generation with lower complexity, data compression, and feature extraction.

To enhance the training stability of GANs, Arjovsky et al. suggested the Wasserstein GAN (WGAN) (Arjovsky, 2017; Arjovsky & Bottou, 2017; Zhang, et al., 2023; Gulrajani, et al., 2017). The WGAN utilizes the Wasserstein distance to calculate the dissimilarity between two distributions. Comparing previous GAN models that primarily relied on conventional KL divergence for optimization, WGANs demonstrate significantly improved performance.

Zhao, et al. introduced the Energy-Based Generative Adversarial Network (EB-GAN), which minimizes energy functions. In this approach, an energy function replaces the discriminator (D) function in the classical generator-discriminator (G-D) framework. The energy function can be calculated through the D function. The EB-GAN aims to generate data that closely resembles the actual data distribution by minimizing the energy function. However, implementing EB-GAN poses certain challenges, including difficulty and computational cost. Therefore, further exploration and improvement are necessary to enhance its practical applications.

Chen, et al. proposed an information-theoretic extension to GAN known as InfoGAN. This extension introduces additional variables, including noise, hidden codes, and variables that maximize mutual information. Consequently, the training complexity and difficulty are increased, particularly for large-scale data and complex applications that require ample training data and computing resources.
Mu Li, et al. introduced Deep Identity-Aware Transfer of Facial Attributes (DIAT) (Li, et al., 2016) in 2016, a facial image conversion technology based on deep learning. The primary objective of DIAT is to transform facial attributes, such as altering facial expressions, age, and hairstyle, while preserving facial identity information. In DIAT, identity features are extracted using the backbone network and fed into the transfer module for further processing. The transfer module aims to retain identity features while transferring attribute features, resulting in changes to facial attributes. Specifically, the transfer module concatenates the encoded identity features with the target attributes and feeds them into a generator to produce the target image. However, DIAT’s performance is heavily dependent on training data. Abundant and high-quality training data are essential to ensure the accuracy and effectiveness of the model.

Casanova, et al. proposed Instance Conditioned GAN (ICGAN) (Casanova, et al., 2021), a variant of GAN that aims to generate images with specific attributes. The key difference between ICGAN and conventional GAN lies in using additional conditions for both the generator and discriminator. These conditions can control the attributes of the produced image, such as object category, color, and pose. In other words, an ICGAN can generate images with desired attributes. The input of an ICGAN consists of two parts: random noise, which contributes to generating image details, and additional conditions that govern the desired attributes of the image. After training, an ICGAN can generate images with specific attributes and produce similar images under previously unseen conditions. However, the diversity of the generated images may be limited, as ICGAN typically excels at generating images with predefined attributes. This limitation restricts the overall diversity of generated images. Therefore, if the objective is to generate images with greater diversity, ICGAN may not be the best choice.

Choi, et al. introduced StarGAN in 2018 and proposed an improved version called StarGAN_V2 (Choi, et al., 2023) at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in 2020. StarGAN is a Generative adversarial network designed for image translation across multiple domains. It enables the transformation of an image into various domains, such as transforming a facial photograph into an image with different hair or eye color, age, and even gender. Although StarGAN can generate images in multiple domains, in practical applications, it may encounter occasional failures or situations where no viable solution is found, resulting in poor quality or inability to generate images.

In 2014, Mehdi Mirza, et al. proposed CGAN (Mirza & Osindero, 2014) based on GAN, which is also considered one of the classic GAN papers. CGAN presents a new method for training and generating network models compared to GAN. It introduces additional conditional input data, denoted as \( y \), which enables simultaneous constraint of both the generator and discriminator. CGAN is a powerful tool for emotional design, as it allows image generation based on specific conditions, such as a particular emotion. For instance, if the desired design aims to evoke a sense of warmth and comfort, the condition can be set as “coziness,” and the CGAN will generate images that convey that particular emotion. By leveraging CGAN, designers can tap into the emotional power of images and create designs that connect with people on a visceral level. A notable application of CGAN in emotional design is the creation of personalized avatars. Avatars represent individuals in virtual worlds, and they can evoke a range of emotions, from excitement to empathy. Through CGAN, designers can generate avatars that reflect the user’s personality and emotions, thereby establishing a deeper emotional connection with the user. However, CGAN encounters challenges related to data scarcity. Sometimes, due to the limited size of the conditional input dataset or the presence of noise or errors in the input conditional data, the generation effect of CGAN can result in significant errors. This paper utilizes a two-stage network for defect complementarity to address this issue.

At the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) International Conference, Isola, et al. from the University of California introduced Image to Image Translation with Conditional Adversarial Networks (pix2pix) (Isola, et al., 2017), which utilized conditional generative adversarial networks to facilitate image to image conversion and grayscale image coloring. This approach requires paired images, where an input image is paired with a corresponding output.
image. For instance, the input image could be a black-and-white sketch of a house, and the desired output image would be the same house in color. Pix2pix takes the input image and generates an output image that aligns with the paired output image. The discriminator in pix2pix evaluates the degree of similarity between the generated and paired output images. Pix2pix is well-suited for tasks requiring precise image-to-image translation, such as colorizing images or transforming sketches into realistic images. However, the pix2pix model necessitates a high-quality dataset. If the dataset’s quality is subpar, the model may encounter issues like overfitting. To address this concern, this article proposes the use of a two-stage network.

A paper titled CycleGAN (Zhu, et al., 2017) presented at the 2017 IEEE International Conference on Computer Vision (ICCV) attracted considerable attention for its application in image style transfer. Unlike conventional generative adversarial networks that typically operate in a single direction, CycleGAN consists of two generators and two discriminators, forming a cyclic structure, as illustrated in Fig. 3. It has also demonstrated impressive results in the field of image enhancement. CycleGAN is another powerful tool for emotional design by enabling image transformations between domains, creating new emotional experiences. For instance, if the objective is to evoke a sense of nostalgia in a design, CycleGAN can transform modern images into vintage ones, evoking a longing for the past. It is worth noting that CycleGAN requires a substantial amount of data.

An example of CycleGAN application in emotional design is the creation of virtual environments. Virtual environments can elicit various emotions, ranging from excitement to relaxation. Using CycleGAN, designers can transform real-world environments into virtual environments that evoke specific emotions, such as a sense of adventure or serenity (Pumarola, et al., 2018).

CycleGAN, on the other hand, is an unsupervised GAN that does not require paired images. Instead, CycleGAN can learn the mapping between two domains even when direct correspondences are unavailable. It employs two generators and two discriminators, each domain having its own set. The generators are trained to translate images from one domain to another and back again, forming a cyclic process. The discriminators assess the similarity between the generated and real images within their respective domains. CycleGAN is particularly well suited for tasks where the image-to-image translation requires flexibility, such as style transfer or image enhancement.

In summary, while both CycleGAN and pix2pix are GANs utilized for image-to-image translation, they differ in their approach and application. Pix2pix requires paired images and excels at precise image-to-image translation tasks, whereas CycleGAN can learn mappings between two domains without the need for paired images and is better suited for flexible image-to-image translation tasks.

However, each of the aforementioned improved GAN models has unique generation characteristics. In the field of emotional design, it is crucial to combine different generation characteristics to achieve the desired emotional design outcomes.

### 3. PROPOSED METHOD

As described in Sections 1 and 2, the GAN has been proven to be one of the best generative models currently available. However, the characteristics of a single GAN network are relatively limited, which may not fully meet the requirements for generating multiple attributes. The two-stage model offers a better solution by addressing the limitations of a single model. It divides the feature learning process into two stages, with each stage carrying out its specific tasks to yield remarkable results. Fig. 1 illustrates the schematic graph of the proposed two-stage framework. The upper part of the figure represents the contour design model. In the input generator network, C is the required conditions and Gaussian noise distribution data, the output is the contour image that reaches the input bar and is input together with the actual image and corresponding labels into the discriminator network, and D is for authenticity identification. The lower part depicts the color design model to receive the contour data obtained from the contour model and perform color design. The two networks are seamlessly coupled to obtain the final result.
The upper half of the image uses cgan to generate contours, while the lower half uses pix2pix for color design.

### 3.1 CGAN for Generating Image Contours

Conditional Generative Adversarial Networks (CGANs) are a type of deep learning algorithm that can generate realistic images based on a given input and a conditional label. One specific application of CGANs is the generation of image contours, which finds utility in various fields, such as graphic design and architecture.

CGAN is a type of GAN that can produce high-quality images with clear contours. It enhances the generation of more accurate and consistent images, leading to improved realism in the output produced by the generator network.

In the context of generating image contours, CGANs can be trained using a dataset comprising images with well-defined contours, such as sketches or line drawings. The conditional label is employed to specify the desired category of object or scene to be generated, such as a building, a landscape, or a person. The generator network takes both the input and the conditional label, producing an image with clear contours that matches the specified category.

One advantage of using CGANs for generating image contours is their ability to produce high-quality images with clear and consistent contours. This quality is particularly valuable in fields like architecture and graphic design, where accuracy and precision in contour depiction are crucial. CGANs can also generate diverse images with different styles and characteristics, depending on the input and the conditional label.

In conclusion, CGANs represent a powerful tool for generating image contours. They can produce high-quality images with clear and consistent contours, which proves useful in various fields, such as architecture and graphic design (Li M, et al., 2020). CGANs offer versatility and flexibility in generating images with well-defined contours by including a conditional label.

The core concept of CGAN is to exercise control over the generated images rather than relying solely on random image generation. Specifically, CGAN incorporates additional conditional information to the inputs of the generator and discriminator. The generator produces images that can only pass the discriminator if they are deemed sufficiently realistic and align with the specified
conditions. In fact, controlling the data generation mode is not feasible in an unconstrained generative model. However, the data generation process can be influenced by imposing conditional constraints and leveraging additional information. Conditional constraints can be class labels, partial image patching data, or even data from different modes. CGANs effectively transform unsupervised learning into supervised learning, enabling the network to learn more effectively under our control.

CGAN is equivalent to adding a condition to both the discriminator and generator parts based on the original GAN. The loss objective function of the original GAN is shown in Equation (1):

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim P_{\text{data}}} \left[ \log D(x) \right] + \mathbb{E}_{z \sim P_{\text{z}}} \left[ \log \left(1 - D(G(z)) \right) \right]$$

where $P_z$ represents the distribution of the data produced by the generator, $P_{\text{data}}$ represents the real data’s distribution, $x$ represents the discriminator’s input, $\hat{x}$ represents the produced output defined as $\hat{x} = G(z)$, with $z$ being the input to the generator $G$ sampled from a pre-defined multivariate distribution, such as the Gaussian distribution used in this paper. The objective of the loss function is to minimize the difference between the generated and real data. Arjovsky, et al. utilized earth-mover distance to train GAN models, commonly known as Wasserstein GAN.

The additional information $Y$ needs to be incorporated into the input of $G$ and $D$. Therefore, the following Equation (2) describes the formulation of the loss function of CGAN:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim P_{\text{data}}} \left[ \log D(x|y) \right] + \mathbb{E}_{z \sim P_{\text{z}}} \left[ \log \left(1 - D(G(z|y)) \right) \right]$$

Figure 2. CGAN Model
The lower part is the generator network, which fuses Gaussian noise and labels and inputs them into the generator. The upper part is the discriminator network, which fuses the training images, labels, and images generated by the generator and inputs them into the discriminator network.

To achieve the purpose of conditional GAN, the principles and training methods of the generator and discriminator networks need to be changed. In the model section, additional label information is included as input for both the discriminator and generator.

3.2 Image to Image Translation for Generating Image Colors

Pix2pix, on the other hand, is a machine-learning algorithm that can generate realistic images from a given input. It falls under the category of generative adversarial networks and uses a neural network to establish a mapping between an input image and an output image. Pix2pix can be utilized for various image-to-image translation tasks, including converting sketches into realistic images, colorizing black and white images, and generating photorealistic images from low-resolution ones.

Pix2pix finds application in emotional design to create visual designs that evoke specific emotions from users. For example, designers can use pix2pix to create a design that evokes a sense of calmness and relaxation. By providing input images representing calmness, such as a serene landscape or individuals meditating, pix2pix can generate designs that reflect those emotions. Similarly, designers can employ pix2pix to produce designs that evoke excitement, joy, or any other desired emotion.

In conclusion, emotional design and pix2pix can be combined to create designs that elicit specific emotions from users. Pix2pix facilitates the creation of visual designs that effectively convey desired emotions, while emotional design enables designers to develop products that establish a strong emotional connection with users. By working together, emotional design and pix2pix can create products that not only have visual appeal, but also evoke positive emotional experiences.

The structure of image-to-image translation bears some similarity to CGAN. First, the generator takes in a real contour or label image, which is used for training. Then, the generated fake image, along with the real contour used as input for the generator, is passed to the discriminator for training. The discriminator is trained using both real images and their corresponding contour images, as depicted in Fig. 3 and Fig. 4. The overall structure is relatively straightforward, and it yields effective results.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In order to demonstrate the effectiveness of the proposed two-stage emotion generation model, a series of experiments were conducted on two datasets: the Radboud Face Dataset (RaFD) database of faces and a self-developed car dataset.

Figure 3. CGAN in the Two-Stage Model
In addition, the generated images from both datasets underwent an assessment of image quality, which was evaluated by utilizing Structural Similarity (SSIM) (Pajovic & Jevremovic, 2023; Park, et al., 2022) and Peak Signal-to-Noise Ratio (PSNR) (Kim & Han, 2022).

PSNR was chosen to measure whether the program’s processing results were satisfactory. It represents the logarithmic value of the mean square error between the processed and the original images relative to \((2^n-1)^2\). The calculation formula is shown in Equation (3):

\[
PSNR = 10 \log_{10} \left( \frac{(2^n - 1)^2}{MSE} \right)
\]  

(3)

The range of PSNR values is (0, inf), and a larger PSNR value indicates a more similar image. The disadvantage of PSNR is that its score may not fully align with the quality perceived by the human eye. On the other hand, the advantage of PSNR lies in its simple algorithm and wide application.

SSIM is a metric used to evaluate the similarity between two images. When there are two images, one being the undistorted image and the other being the distorted one, the structural similarity index indicates the distorted image quality. In contrast to the peak signal-to-noise ratio, the structural similarity index measures image quality that is more in line with human perception. Using signals \(x\) and \(y\) as examples, the calculation formula for structural similarity is shown in Equation (4):

\[
SSIM(x, y) = \left( \frac{1(x, y)}{c(x, y)} \right) \left( \frac{1(x, y)}{s(x, y)} \right) \left( \frac{1(x, y)}{c(x, y)} \right)
\]

(4)

\[
1(x, y) = \frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}
\]

(5)

\[
c(x, y) = \frac{2\delta_x \delta_y + C_2}{\delta_x^2 + \delta_y^2 + C_2}
\]

(6)
where \( l(x, y) \) compares the brightness between two signals, \( c(x, y) \) compares the contrast between the two signals, and \( s(x, y) \) compares the structure of the two signals. The parameters \( \alpha > 0, \beta > 0, \) and \( \gamma > 0 \) are primarily used to adjust the importance of \( l(x, y), c(x, y), \) and \( s(x, y) \), while \( \mu_x \) and \( \mu_y \) represent the average values of the signals, \( \delta_x \) and \( \delta_y \) denote their standard deviations, and \( \delta_{xy} \) represents their covariance. Constants \( C_1, C_2, \) and \( C_3 \) are used to maintain the stability of \( l(x, y), c(x, y), \) and \( s(x, y) \). A larger value of SSIM indicates a higher similarity between the two signals. When calculating the similarity of image structures, we typically set \( \alpha = \beta = \gamma = 1 \) and \( C_3 = C_2/2 \), resulting in the final formula for SSIM.

\[
SSIM(x, y) = \frac{(2\mu_x\mu_y + C_3)(2\delta_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\delta_x^2 + \delta_y^2 + C_2)}
\]

4.1 Experiments on RaFD

The RaFD dataset (Langner, et al., 2010; Yin, et al., 2022) was created by the Nijmegen Institute for Behavioral Sciences at the University of Radbold in Nijmegen, Netherlands. It consists of 67 models’ images depicting eight emotional expressions. The dataset includes images of Caucasian males and females, Caucasian children (boys and girls), and Moroccan Dutch males, as illustrated in Fig. 5.

RaFD is a high-quality facial database with images of individuals displaying eight emotional expressions. These expressions are based on the facial action encoding system including disgust, anger, fear, sadness, happiness, surprise, neutrality, and contempt. Each emotion is captured from three different gaze directions, and the images are simultaneously taken from five different camera angles. A partial dataset is shown in Fig. 4.

Because this dataset does not provide facial contour data, we extracted the contours of the images ourselves. As depicted in Fig. 6, we used an algorithm to convert the color images into outline images.

Figure 5. RaFD dataset which shows six different expressions of four people
Fig. 6 displays the transformed process of converting the color pictures into sketch images, as described below:

1. Convert the color images into grayscale images, then perform image inversion on the grayscale images.
2. Apply Gaussian blurring to the grayscale images to reduce noise.
3. Fuse each gray image with the corresponding Gaussian-blurred image to obtain the sketch image.

Some conversion results are listed in Fig. 7.

After preparing the dataset, we will compare the two-stage model with the Conditional GAN. The results of this comparison are shown in Fig. 8. The two-stage model generates two sets of data: the contour data in the first row and the color data in the second row. CGAN generates the first row of data, and the second is generated by pix2pix. Similarly, the third row of data is generated using the CGAN single model. When comparing the data in the second and third rows, it can be observed that the two-stage model produces superior results.

The range of SSIM is [0,1], and a higher value indicates better image quality. When two images are identical, the SSIM value is 1. Therefore, the SSIM values between the generated results and the corresponding expressions in the RaFD dataset were calculated to evaluate the experiment.

Figure 6. Image contour extraction image

Figure 7. Examples of original images and their sketch maps in the RaFD dataset
results are presented in Table 1. Comparing the SSIM values for the eight different expressions, the average result improved by 31.63%. In comparison with the results of the single-stage model in four references, all the comparison results showed improvement. This comparison demonstrates the full effectiveness of the multi-stage model.

We also plotted the evaluation result histogram as shown in Fig. 9, and the evaluation results of all 9 expressions improved, fully proving the effectiveness of the two-stage model.

The comparative results of PSNR are presented in Table 2. The two-stage model proposed in this paper outperforms the CGAN model by 9.8%. The corresponding histogram of the results, as illustrated in Fig. 10, further confirms that our model outperforms other single-stage models.

### 4.2 Experiments on Car Images

This article conducted a search for 3000 car images on the internet for testing purposes and selected 1000 sports car, car, and SUV images with relatively simple backgrounds as the self-developed images for this paper, as shown in Fig. 11.

<table>
<thead>
<tr>
<th>Emotional</th>
<th>Two-stage model</th>
<th>CGAN (Mirza M. et al., 2014)</th>
<th>DIAT (Li M. et al., 2016)</th>
<th>IcGAN (Casanova A. et al., 2021)</th>
<th>StarGAN (Choi Y. et al., 2023)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry</td>
<td>0.650</td>
<td>0.603</td>
<td>0.449</td>
<td>0.661</td>
<td>0.663</td>
</tr>
<tr>
<td>Contemptuous</td>
<td>0.673</td>
<td>0.503</td>
<td>0.435</td>
<td>0.67</td>
<td>0.642</td>
</tr>
<tr>
<td>Disgusted</td>
<td>0.651</td>
<td>0.475</td>
<td>0.415</td>
<td>0.602</td>
<td>0.667</td>
</tr>
<tr>
<td>Fearful</td>
<td>0.663</td>
<td>0.522</td>
<td>0.432</td>
<td>0.569</td>
<td>0.689</td>
</tr>
<tr>
<td>Happy</td>
<td>0.749</td>
<td>0.461</td>
<td>0.432</td>
<td>0.536</td>
<td>0.684</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.692</td>
<td>0.516</td>
<td>0.454</td>
<td>0.595</td>
<td>0.643</td>
</tr>
<tr>
<td>Sad</td>
<td>0.634</td>
<td>0.529</td>
<td>0.438</td>
<td>0.517</td>
<td>0.622</td>
</tr>
<tr>
<td>Surprised</td>
<td>0.650</td>
<td>0.464</td>
<td>0.436</td>
<td>0.572</td>
<td>0.637</td>
</tr>
<tr>
<td>Average</td>
<td>0.670</td>
<td>0.509</td>
<td>0.436</td>
<td>0.59</td>
<td>0.656</td>
</tr>
</tbody>
</table>
We will also convert all the images into grayscale outline images, as shown in Fig. 12. The generation results of the two models are also compared, as shown in Fig. 13. The left side illustrates the results generated by a single model, while the right side illustrates the results generated by the dual model. The final results show that the results generated by the two-stage model are significantly superior to those generated by a single-stage model.

Tables 3-5 compare the results of sports cars, cars, and SUVs, using SSIM and PSNR as performance indicators. According to the results in the tables, the results of this article are superior to those of CGAN.
Table 3 compares the PSNR and SSIM evaluations of 10 randomly selected images of sports cars. The comparison shows that the two-stage model in this article performs better in both evaluation indicators than the single-stage model, with a 41% increase in PSNR and a 36.5% increase in SSIM indicators.
Figure 12. Image contour extraction

Figure 13. Comparison results on dataset car images

Table 3. Evaluation and comparison of the sports cars’ results with PSNR and SSIM

<table>
<thead>
<tr>
<th>image_id</th>
<th>Two-stage emotion generation model</th>
<th>Conditional GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>1</td>
<td>7.120</td>
<td>0.540</td>
</tr>
<tr>
<td>2</td>
<td>11.593</td>
<td>0.875</td>
</tr>
<tr>
<td>3</td>
<td>8.843</td>
<td>0.706</td>
</tr>
<tr>
<td>4</td>
<td>7.043</td>
<td>0.532</td>
</tr>
<tr>
<td>5</td>
<td>10.129</td>
<td>0.762</td>
</tr>
<tr>
<td>6</td>
<td>9.928</td>
<td>0.752</td>
</tr>
<tr>
<td>7</td>
<td>11.406</td>
<td>0.863</td>
</tr>
<tr>
<td>8</td>
<td>5.221</td>
<td>0.383</td>
</tr>
<tr>
<td>9</td>
<td>9.272</td>
<td>0.749</td>
</tr>
<tr>
<td>10</td>
<td>11.633</td>
<td>0.935</td>
</tr>
<tr>
<td>Average</td>
<td>9.219</td>
<td>0.710</td>
</tr>
</tbody>
</table>
Similarly, Table 4 extracts 10 sets of images of cars for evaluation and comparison between PSNR and SSIM. The PSNR index increased by 24%, and the SSIM index increased by 20%.

Finally, Table 5 randomly selected 10 sets of images from the SUV to evaluate and compare PSNR and SSIM. The PSNR index increased by 10.5%, and the SSIM index increased by 4.2%.

Through experiments, it can be observed that the two-stage model has a significantly better effect on image generation than the single-stage model. The two-stage model adopts a step-by-step generation strategy, which allows for decomposing complex tasks and ultimately achieves good results.

Table 4. Evaluation and comparison of the cars' results with PSNR and SSIM

<table>
<thead>
<tr>
<th>image_id</th>
<th>Two-stage emotion generation model</th>
<th>Conditional GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>1</td>
<td>7.764</td>
<td>0.591</td>
</tr>
<tr>
<td>2</td>
<td>12.270</td>
<td>0.931</td>
</tr>
<tr>
<td>3</td>
<td>9.583</td>
<td>0.762</td>
</tr>
<tr>
<td>4</td>
<td>9.305</td>
<td>0.705</td>
</tr>
<tr>
<td>5</td>
<td>12.359</td>
<td>0.931</td>
</tr>
<tr>
<td>6</td>
<td>10.707</td>
<td>0.820</td>
</tr>
<tr>
<td>7</td>
<td>8.450</td>
<td>0.641</td>
</tr>
<tr>
<td>8</td>
<td>7.471</td>
<td>0.551</td>
</tr>
<tr>
<td>9</td>
<td>10.243</td>
<td>0.833</td>
</tr>
<tr>
<td>10</td>
<td>10.322</td>
<td>0.829</td>
</tr>
<tr>
<td>Average</td>
<td>9.847</td>
<td>0.759</td>
</tr>
</tbody>
</table>

Table 5. Evaluation and comparison of the SUV's results with PSNR and SSIM

<table>
<thead>
<tr>
<th>image_id</th>
<th>Two-stage emotion generation model</th>
<th>Conditional GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>1</td>
<td>9.247</td>
<td>0.702</td>
</tr>
<tr>
<td>2</td>
<td>9.973</td>
<td>0.752</td>
</tr>
<tr>
<td>3</td>
<td>7.773</td>
<td>0.606</td>
</tr>
<tr>
<td>4</td>
<td>5.983</td>
<td>0.450</td>
</tr>
<tr>
<td>5</td>
<td>8.989</td>
<td>0.677</td>
</tr>
<tr>
<td>6</td>
<td>10.094</td>
<td>0.835</td>
</tr>
<tr>
<td>7</td>
<td>9.637</td>
<td>0.733</td>
</tr>
<tr>
<td>8</td>
<td>4.799</td>
<td>0.350</td>
</tr>
<tr>
<td>9</td>
<td>8.150</td>
<td>0.650</td>
</tr>
<tr>
<td>10</td>
<td>9.312</td>
<td>0.744</td>
</tr>
<tr>
<td>average</td>
<td>8.396</td>
<td>0.650</td>
</tr>
</tbody>
</table>
5. CONCLUSION

This paper proposes a two-stage emotional image generation model that combines CGAN and Pix2Pix. The main advantage of the proposed model is the use of two separate networks, one for contour and another one for color, to facilitate feature learning. These networks operate independently without interfering with each other’s learning. The two-stage design model in this paper can generate two types of design drawings that can inspire designers. The model’s effectiveness is validated through both SSIM and PSNR evaluation metrics. The RaFD dataset increased by 31.63% and 10.78%, respectively, while the automotive dataset increased by 25% and 20%, respectively.

Because this article only explores two design elements, there is a limitation of this research and other design elements cannot be added without more computational costs. Future work will incorporate additional elements, such as scenarios and dynamic effects. The research findings are anticipated to inspire more researchers to develop models with increased information capacity and provide successful methods for other application areas in the long term.

AUTHOR NOTE

The authors of this publication declare there are no competing interests. The data used to support the findings of this study are included within the article. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the author(s) of the article.
REFERENCES


