Visual-attention GAN for interior sketch colourisation

Xinrong Li 1 | Hong Li 1 | Chiyu Wang 1 | Xun Hu 1 | Wei Zhang 2

1 School of Computer Science and Engineering, Central South University, Changsha 410000, China
2 School of Architecture, Hunan University, Changsha 410000, China

Abstract
In the professional field of interior designing, sketch colouring is often a time-consuming and rapidity task. The traditional neural network does not handle the semantic relationship of sketch lines well, and the colouring effect is unsatisfactory. This paper proposes visual-attention generative adversarial network (VAGAN), which enhances the processing effect of edge semantics, strengthens the network to line edge recognition ability, as well as reduces colour overflow and improved model colouring result. In addition, a two-stage training mode is used to simplify the training of rare samples. The simple line draft input into the trained VAGAN, output natural, realistic colour pictures. The experimental results show that, compared with the existing methods, the proposed method can better deal with the problem of sketch and generate stable and reliable images.

1 | INTRODUCTION

Sketch exists in the fields of pattern design, animation creation and video editing, a vivid and characteristic colour picture can depict people’s desire for colour. However, completing the task of colour combination and line painting is not easy, it requires the designer's imagination to matching the colour of the space objects and depicting the light and shadow details in the image [1]. In the traditional field of interior design, create often starts with a few understated strokes, the original works only contain the outline of the edge of the object, not the texture features of details [2], due to their unlimited subjects and texture styles, after finishing the first draft, designers spend a lot of time colouring sketches. An automated colouring system can dramatically reduce designer creation time and increase their productivity.

Neural network algorithms for style transfer and greyscale image colouring [3], [4], which combine content and style maps to produce stunning and near-perfect images. For interior design sketch colouring tasks, the process of colouring sketch or line art is not like the traditional greyscale picture, because greyscale pictures contain greyscale information of images and object texture information, neural network can be combined with local image features and global priors of convolution calculations to achieve accurate image colouring. However, the lack of texture information and unclear edge information brings a lot of difficulties in sketch colouring; neural network has to learn a lot of high-level image information implicitly, including the edge information of objects and the semantic information of objects themselves. For example, the colouring range cannot exceed the semantic area of the single object [5], the neural network must be able to identify the semantic segmentation of multiple objects [6] and can be colour for specific objects, including the colour of wooden floors is yellow, the colour of plants is green, the ceiling lights will have obvious highlights and so on.

In this article, we propose a full-automatic method that generates interior design images with colour and texture. We divide the sketch colouring process into two-stage tasks, the first stage is to solve the problem of entity recognition in sketching. In real-world pictures, there is an obvious greyscale transition between objects. Based on this information, the entity segmentation of real-world images becomes an easier task. For line drafts, the lack of obvious border transitions as well as the lack of texture semantic information makes the entity recognition less accurate and colouring staining. We note that the visual-attention algorithm is helpful for edge information extraction, model using this algorithm can better handle semantic differences between different objects in the face of complex spatial environments and retain more semantic information, the process of this part improves the accuracy of subsequent colouring. We implement image to image
translation in the second stage, adopting a conditional generative adversarial network with cycle-consistent structure. Network training does not require matching datasets that provides unparalleled convenience for dataset acquisition. In addition, our training process is multi-stage, from partial transition to overall two-stage training which increases the stability of sketch colourisation.

We validate our approach on a set of interior sketches with a wide range of content and styles. We also compare our results in multiple experiments. Statistics show that our method outperforms existing ones in terms of colour quality and convenience. Our contributions can be summarised as follows:

- We propose an automatic colouring model in the field of interior line drawing, which can learn to generate realistic images from sketches with missing textures.
- We improve the quality of sketch-to-image synthesis compared to existing work [7]. We proved the feasibility of the method through the control experiment.
- We show that our method can generate different styles of interior sketch according to different styles of real-world reference images. We achieve this generality by expanding training data using different interior styles.

2 RELATED WORK

Image synthesis has always been a hot issue in the field of computer vision. In the past, with insufficient computing power, the main method of image generation was traditional machine learning. Based on data driving or instance mapping, they have a good performance in dealing with style transfer and grey image colouring, but these methods are often driven by a single data, image generation often takes a lot of time, and in the face of complexly distributed scenes, it is difficult to find a suitable expression map, and the results are often unsatisfactory. The success of deep convolutional network in image classification has made neural network gradually applied in the field of image synthesis. The subsequent generative adversarial network greatly enriches the application field of image synthesis. Since then, various methods have emerged in an endless stream.

2.1 Traditional machine method

Traditional machine methods for solving image colouring problems can be divided into two types: parametric modelling method based on statistical distribution [8–10] and non-parametric modelling method based on Markov field [4], [11–13]. In the past, due to the lack of computing power, the main methods were non-parametric methods. Based on data-driven and instance mapping, they have excellent performance in dealing with style transfer and greyscale image colouring [14–16]. However, these methods are often driven by a single data, image generation often takes a lot of time, and in the face of complex distributed scenes, it is difficult for the model to find a suitable expression mapping, and the results are often unsatisfactory.

2.2 Convolutional neural network

In recent years, parametric methods have gradually become mainstream, significantly different from traditional non-parametric methods, these methods have very large computational models and tens of thousands of image training datasets, so they have better performance in image synthesis. The success of the deep convolutional network [17] has made the development of the image field more diverse. Gatys et al. [3] first studied the use of the neural network to solve the problem of image style transfer, making it possible to separate and reorganise the content and style of images. Subsequently, neural networks for style transfer have been adopted in [18–20]. Iizuka et al. [21] proposed a new technique for automatic colouration of greyscale images combined with global prior and local image features. Yao et al. [22] develop an attention-aware multi-stroke style transfer model to coordinate the spatial distribution of visual attention between the content image and stylised image. Park et al. [23] integrate the image of local style patterns and realised the transformation of any style through several style reference diagrams. These parameterisation methods combined with local features to solve the problem of image translation well.

2.3 Generative adversarial networks

Generative adversarial network (GAN) is a hot area of artificial intelligence. Proposed by Goodfellow et al. [24] in 2014, the model consists of a generator and a discriminator that generates data in a way that it is adversarial learning. Radford et al. [25] combined GAN with CNN for the first time to improve the learning performance of the model by utilising the powerful feature extraction ability of CNN. Mirza et al. [5] introduced conditional constraints on the basis of GAN, which solved the disadvantage that the generated data is uncontrollable to some extent. Pix2pix [6] is a conditional adversarial network which has a good performance in colouring simple object, but it is not satisfactory in complex scenes. Hensman et al. [26] require only a single colourised reference image for comic colouring, avoiding the need for a large dataset. A two-stage generative adversarial network model [27], [28] is used to colour and adjust the colours of the line drawing cartoon character and render different levels of detail with different brush strokes, create images that rival those of professional painters. Scribble [1] utilised conditional feedforward convolutional network to colour the sketches and could manually specify the colouring colours of specific targets, generate realistic cars, bedrooms, or faces, etc.
2.4 Visual attention mechanism

Visual attention in deep learning guides network to focus on regions of interest relevant for a particular recognition task, which avoids computing features from irrelevant image regions and resulting in better performance. Tang et al. [29] proposed a multi-channel attention method based on scene image and new semantic mapping to solve cross-view image translation. Sudhakaran et al. [30] use attention help to identify human actions from video, surpassing the performance of non-attentive alternatives. A kind of attention network based on pyramid features for salient object detection has been adopted in [31–33], improving the accuracy of target detection in complex scenes. Human visual perception shows good consistency in the recognition of distorted images, according to the characteristics of human attention, Guo [34] uses attention heat map to improve the robustness of image classification. Guo et al. [35] and Fu et al. [36] use self-attention to solve scene segmentation task which segments foreground objects at the instance level as well as background contents at the semantic level. Visual attention mechanism is widely used in image segmentation, object detection, video prediction and other image fields.

2.5 Current work in the sketch

The application of sketch is not only in the field of colouring, but recent research also shows that sketch is widely used in image retrieval, motion research, image restoration, etc. Muhammad et al. [37] train a model for abstract sketch generation through reinforcement learning of a stroke removal policy that learns to predict which strokes can be safely removed without affecting recognisability, in addition, the author also proposes the concept of image retrieval based on a fine-grained sketch. With the popularity of stroke learning, a cross-modal binary representation learning method called zero-shot learning [38–40] has been proposed. This way of hand-drawn search images by users solves the cross-modal image retrieval of sketches to real-world pictures to some extent. Dutta and Akata [41] and Wan et al. [42] bring data matching in different fields to face recognition, enabling face sketch recognition from sketch to photo. Since there may be no lines from the photo to the sketch that is perfectly aligned with the image features, Yi et al. [43] developed a novel loss to measure the similarity between generated sketch and artists drawings based on distance transforms and realised the generation of an artistic portrait from the face photo. Chen and Hays [44] demonstrated a data augmentation technique for sketches that synthesising realistic images from human-drawn sketches.

3 VISUAL-ATTENTION GENERATIVE ADVERSARIAL NETWORK

The proposed visual-attention generative adversarial network (VAGAN) is a feedforward convolutional network essentially. The network depicts the missing colour details of the line art by learning the mapping between the input image and the ground-truth image. The visual attention algorithm enables the deep generation network to retain more fine-grained information and strengthen the network’s ability to recognise edge information. The network structure is end-to-end and cycle-consistent. The end-to-end network structure eliminates the manual data pre-processing, and the cycle-consistent training method does not require matching datasets, which reduces the difficulty for the user to obtain data. In addition, we have improved the loss function, and the network can sprinkle colours more accurately. The relevant results are shown in Figure 1.

3.1 Visual attention generator

Visual context is very important for image generation. The fully connected neural network is not suitable for our current work because there are too many parameters and a weak correlation between pixels. Current CNN models learn image features by stacking multiple convolutional layers and pooling layers. Using bottom-up convolutional and pooling layer directly may not be able to effectively process complex images. In the sketch colourisation, the network needs to handle the mapping relationship between sparse lines and ground truth. The convolution kernel must capture a large enough acceptance domain to meet the above conditions. To this end, we propose a visual-attention generator (VAG), as shown in Figure 2.

VAG is a fully convolutional neural network. The network consists of an encoder and a decoder. The encoder downsamples the image to obtain an edge feature. The decoder recovers position details by upsampling. However, upsampling will cause edge blurring and loss of position details. Some existing work uses skip connections to connect low-level features to high-level features, which helps supplement location details. But due to the lack of bottom semantic information, it contains a lot of useless background information. This information may interfere with the segmentation of the target object. To solve this problem, VAG designed a visual attention module (VAM) based on the decoder recovery feature to capture high-level semantic information and emphasise target feature. The enhanced attention module is placed in the upsampling process. We use the eigenvalues of conv3, conv7 and conv10 to calculate the attention value. In addition, the convolution kernel size of the network is $3 \times 3$, pooling is performed using max-pooling, the step size is 2, and the decoder uses transposed convolution for upsampling to obtain accurate edges.

3.2 Visual attention module

Different convolution kernels correspond to a specific semantic response, and the interior scene often involves multiple semantic information. Therefore, the VAM emphasises target semantic through semantic dependence. It captures semantic information in the high-level feature map and global context in the
FIGURE 1 With visual-attention GAN, one can automatically transform the interior sketch into colourful interior pictures. Top row: interior sketches. Bottom row: results from VAGAN. GAN, Generative adversarial network; VAGAN, Visual-attention generative adversarial network

FIGURE 2 The architecture of the visual attention generator. The generator consists of an encoder and a decoder. The encoder is used to capture deep semantic features. The decoder consists of a visual attention module and a transposed convolution

low-level feature map to encode semantic dependence. The high-level feature map contains rich semantic information and can be used to guide the low-level feature map to choose important location details. In addition, the global context of the low-level feature map encodes the semantic relationship between different channels, which helps to filter the interference information. VAM effectively uses this information to highlight target areas and improve feature representation. VAM is shown in Figure 3.

VAM first adds a low-level feature map to a high-level feature map to get the feature map after attention. Compared with stitching [45], addition can reduce convolution parameters and help reduce computational costs. The vectors are then batch processed for standardised convolution to further capture semantic dependencies. VAM uses the Softmax and ReLU as the activation function to normalise the vector. Finally, the resampler is used to extract the global context and semantic information. As described in Equation (3), global information is compressed into a focus vector, and the semantic dependencies are encoded, which helps to emphasise key features and filter background information. In addition, since a $1 \times 1$ convolution kernel is used, this module does not add too many parameters, which also greatly reduces the calculation cost. The vector value generation process is as follow:

$$\alpha (x_k) = R \left( f_r (\omega (f_r (g(x_k)) + f(x_k)) + \beta) \right)$$ (1)

$$y(x_k) = \alpha (x_k) + f(x_k)$$ (2)

where $x_k$ represents the convolution kernel of the current layer network, formula as $H_k \times W_k \times D_k$, which indicates the size and number of the convolution kernel. $g(x_k)$ and $f(x_k)$ are high-level and low-level feature map, respectively. $f_r(x_k)$ refers to the ReLU function, and $f_r(x_k)$ refers to the Softmax function. $\omega(x_k) + \beta$ is a convolution kernel with a parameter of $1 \times 1$. $\beta$ is the bias of the convolution kernel. $R(x_k)$ represents the
resampler

\[ R (x_k) = \frac{1}{WXH} \sum_{i=1}^{W} \sum_{j=1}^{H} x_{k}(i,j) \] (3)

where \( k = 1, 2, ..., c \) and \( x = [x_1, x_2, ..., x_c] \).

3.3 | Cycle process

The essence of sketch colouring is the problem of image translation. In the past, training GAN required paired datasets. Taking advantage of the fact that CycleGAN [7] does not require a matching dataset, VAGAN is proposed. The schematic drawing process of the network colouring process is shown in Figure 4.

Network training requires two types of datasets: dataset \( a \sim p_{\text{data}} (A) \) and dataset \( b \sim p_{\text{data}} (B) \), where \( a \sim p_{\text{data}} (A) \) is the sketches and \( b \sim p_{\text{data}} (B) \) is the ground truth. The cycle process is composed of double generators and double discriminators, both of which are full convolutional network. The goal of VAGAN is to learn the mapping between training samples \( \{a_i\}_{i=1}^{N} (a_i \in A) \) and \( \{b_i\}_{i=1}^{N} (b_i \in B) \). GeneratorA2B converts the sketch into a fake ground-truth picture FakeB. The role of FakeB is to enter DiscriminatorB together with the ground-truth and calculate the first set of adversarial loss. Subsequently, the ground-truth will be generated by GeneratorB2A with a fake sketch FakeA, and FakeA will calculate the second set of adversarial loss together with the original sketch.

Unimproved GAN networks are often composed of a single generator and discriminator. They need paired datasets, so only
a set of adversarial loss need to be calculated to achieve a better image transfer effect. For training on unpaired datasets, GAN cannot implement image remapping well, so additional cycle-consist loss needs to be added. We send FakeB and FakeA to GeneratorB2A and GeneratorA2B, respectively, and we can get the restored style pictures CycleA and CycleB. These pictures and their corresponding style discriminators will work together to calculate the cycle-consist loss. The introduction of cycle-consist loss reduces the possibility of mapping paths and allows GAN to implement the mapping between images in the simplest way (maintaining the image structure).

VAGAN can handle the mapping between sketches and ground-truth without the paired datasets. The principle is to use CycleGAN’s special ring structure to cyclically transform the images in the two datasets to achieve style conversion between different data. At this stage, due to the addition of VAM to the generator, the deep network model can learn the global colour mapping of the sketch and predict the corresponding colour sketch more accurately.

### 3.4 Loss function

In the sketch colourisation, the loss function represents the price paid for inaccurate colouring prediction. The loss function in VAGAN mainly includes two parts: adversarial loss and cycle-consist loss, as follow:

\[
L(G_{A2B}, G_{B2A}, D_{A}, D_{B}) = L_{GAN}(G_{A2B}, D_{B}, A, B) + \lambda L_{cyc}(G_{A2B}, G_{B2A})
\]

(4)

\[
L_{GAN}(G_{A2B}, D_{B}, A, B) = \mathbb{E}_{a \sim p_{data}(a)} \left[ \| \log D_{B}(a) \|_1 \right]
\]

(5)

\[
+ \mathbb{E}_{b \sim p_{data}(b)} \left[ \| \log (1 - D_{B}(b)) (G_{A2B}(b)) \|_1 \right]
\]

where \( L_{GAN}(G_{B2A}, D_{A}, X, Y) \) and \( L_{GAN}(G_{A2B}, D_{B}, X, Y) \) are the adversarial loss of the bilateral mapping; \( \lambda L_{cyc}(G_{A2B}, G_{B2A}) \) denotes the cycle-consist loss, and \( \lambda \) represents the loss coefficient, which controls the importance of the cycle-consist loss relative to the adversarial loss, \( \lambda L_{cyc}(G_{A2B}, G_{B2A}) \) is the \( L1 \) norm between the input image and the reconstructed image [46]. The cycle-consist loss is mainly composed of two parts: the \( L1 \) loss between CycleA and the original sketch, and the \( L1 \) loss between CycleB and ground-truth. The formula is shown as:

\[
L_{cyc}(G_{A2B}, G_{B2A}) = \mathbb{E}_{a \sim p_{data}(a)} \left[ \| CycleA - a \|_1 \right]
\]

(6)

\[
+ \mathbb{E}_{b \sim p_{data}(b)} \left[ \| CycleB - b \|_1 \right]
\]

The purpose of cycle-consist loss is to preserve the input contour information and capture features of the target domain, such as texture, colour or style. The training process of the model can also be summarised as the following optimisation problems:

\[
G^*, F^* = \arg \min_{G_{A2B}, G_{B2A}, D_{A}, D_{B}} \max L(G_{A2B}, G_{B2A}, D_{A}, D_{B}).
\]

(7)

The above formula represents a dynamic game between GAN. The generator needs to make the pictures it produces realistic enough to improve the recognition score, and the discriminator needs to improve its recognition ability to reduce the recognition score. This game is the opposite of mathematical expression. When the generator and discriminator reach the best performance, the game finally reaches equilibrium.

Our goal is to generate realistic interior colouring pictures. The loss of the original adversarial loss does not accomplish the above tasks well. As described in [2], it is not enough to reproduce the fine-grained texture details with the original loss. Sketch colouring is a delicate issue. It is not only the image mapping problem but also the colour overflow and shadow control. Therefore, on the basis of the original, we have improved the adversarial loss function and proposed \( L_{VAGAN}(G_{A2B}, D_{B}, X, Y) \) to make it perform better in processing texture structure and colour expression:

\[
L_{VAGAN}(G_{A2B}, D_{B}, A, B) = \mathbb{E}_{a \sim p_{data}(a)} \left[ (D_{B}(a) - 1)^2 \right] + \mathbb{E}_{b \sim p_{data}(b)} \left[ D_{B}(G_{A2B}(b))^2 \right].
\]

(8)

The loss function of the original GAN often uses the \( s \)-shaped cross-entropy loss function. However, this loss function may cause the gradient to disappear during the learning process. To overcome this difficulty, we use the least-square loss instead of the cross-entropy loss, which has two advantages over the original GAN: VAGAN can generate higher quality images, and the learning process is more stable.

### 4 TRAINING PROCESSING

This article adopts the characteristics of transfer learning [47], using different types of datasets to train network (including chairs, beds, indoors, etc.) at different stages. The training process is simple to complex. In the early stage, a simple type of dataset is used. Such a dataset contains independent objects of different styles, shapes and patterns. In the later stage, the non-matching dataset in complex scenes is employed. The complex dataset mainly consists of interior real pictures and interior hand-drawn line drafts. This section will introduce the acquisition and processing methods of the dataset. We will also provide detailed environment configuration and parameters for the process of transfer learning.

#### 4.1 Datasets collection

Dataset acquisition mainly utilises crawlers to collect a large number of real images on the Internet, a total of 2200 sheets. Considering the complexity of interior objects, we collected different types of datasets for different objects, including matching and non-matching data of individual individuals (a total of 500 pairs, ratio 3 to 2) and matching and non-matching data of the overall interior pattern (a total of 600 pairs, ratio 1 to 5). Some
4.2 Data processing

In the data processing, we randomly clip the dataset, rotate the image, adjust the contrast and brightness information of the line image and scale the image to achieve the same object with different spatial sizes to improve the richness of the dataset.

We expect the network to understand the semantic information of different objects (for example, the light bulb will shine, the carpet is dark, the ground is brown), and the network needs enough data to meet the learning needs of different features. Due to the lack of interior sketches, it is difficult to use all interior sketch pictures to train models and use the input sample to correspond to a specific image area to achieve the colouring of the line draft. In this article, the training is divided into two stages: feature learning and style learning. Feature learning learns the characteristics of local objects, and style learning learns the overall style characteristics.

4.3 Training setup

The deep learning framework is Pytorch [49], which uses Adam [50] algorithm as the learning rate optimiser, the initial learning rate is 1e-2, and the network’s batch size is 32. The epoch of the stage is 500 and the second stage is 1000. The model uses NVIDIA GeForce TITAN throughout the training and testing process. Our testing process is fully automated and automated processing automatically crops the sample into patches, so the model also supports image input at any resolution.
EXPERIMENTAL RESULTS AND DISCUSSIONS

5.1 Performance of our method

In this section, we compare the shading capabilities of the model under different conditions. The effects of the loss function and VAM on the colouring results were studied. We train it using the same samples of interior sketches, keeping the other variables the same. The test results are shown in Figure 8.

We compared the colouring performance of CycleGAN, CycleGAN + least-squares loss, CycleGAN + VAM and VAGAN. It can be observed from Figure 8 that in the pictures generated by CycleGAN, the model cannot accurately distinguish the relationship between the local and the whole in the image. Due to the unequal mapping information, the decoder cannot rely on the position information provided by the encoder. Local objects are coloured based on global semantics, and the model is easily disturbed by deep features, which is catastrophic for the model’s colouring results, easily causing colour overflow and inaccurate colouring. We also tested the effect of the loss function on the model. Changing the s-shaped cross-entropy loss to least-squares loss does not substantially improve the image generation quality of CycleGAN. Intuitively, the sharpness of the contours of local objects is improved to a certain extent. During the training process, the model’s convergence time becomes shorter.

The addition of VAM is the most important factor affecting image generation. CycleGAN and VAGAN with VAM are similar in the final colouring performance. Both of them can restore the approximate real scene according to the context mapping information provided by VAM. Specifically, VAM clearly highlights the correct target area, well suppressed the significance of the background area. With the edge retention loss, a prominent map with clear boundaries and consistent saliency is generated. In addition, VAGAN implements the two methods described above. VAM captures the spatial changes, light projections and other details that exist in real-world pictures and reflects them in colour pictures to achieve natural colouring that matches the characteristics of reality. The addition of least-squares loss speeds up the model’s convergence speed and generates a more stable coloured picture.

In order to evaluate the performance of the colouring effect, we used the standard metrics of the Cityscapes benchmark [51], including per-pixel accuracy, per-class accuracy and average class intersection-crossing (class IOU). In Table 1, removing the VAM will greatly reduce the qualities, and removing the least-squares loss will also indirectly affect the results. There-

![Figure 8](https://example.com/figure8.png)

**Figure 8** Different methods for mapping sketch images. From left to right: input, CycleGAN, CycleGAN + least-squares loss, CycleGAN + VAM, VAGAN (ours). VAM, Visual attention module; VAGAN, Visual-attention generative adversarial network.
TABLE 1  Classification performance of the model for different losses, evaluated on sketches

| Loss                | Per-pixel acc. | Per-class acc. | Class IOU |
|---------------------|----------------|----------------|-----------|
| CycleGAN            | 0.52           | 0.17           | 0.11      |
| CycleGAN + least-squares | 0.53           | 0.18           | 0.12      |
| CycleGAN + VAM      | 0.56           | 0.20           | 0.12      |
| VAGAN (ours)        | 0.57           | 0.21           | 0.13      |

Abbreviations: VAM, Visual attention module; VAGAN, Visual-attention generative adversarial network.

FIGURE 9  The effect of different training samples on the results

Therefore, we concluded that both of these were crucial to our final colouring results.

5.2  Comparison with different training methods

Keeping other settings the same, we train the network using different training methods to analyse how they influence the resulting quality. The dataset of the individual samples is used for training in the first training mode. In addition, interior art pictures are used for training in the second training mode. Finally, the two datasets are combined for training. The effect of different datasets on the generated image is shown in Figure 9.

In Figure 9, if only the first training samples are used to colour the interior sketch (results in the second column), the model can accurately grasp the individual’s semantic information and the corresponding edges, and effectively reduce the colour overflow. However, since there is no light and shadow information of the real-world interior picture, the coloured picture does not reflect the real-world light changes, nor does it consider the specific colour of a particular area. If the second training samples (results in the third column) are adopted, it is difficult for the model to convey the texture information to a single individual, and the individual colour will have a slight edge overflow, and in a few scenes, the picture is not coloured. Finally, we adopted the training samples of the joint dataset (results in the fourth column) to solve the problem in the above training mode to a certain extent.

5.3  Comparison with different datasets

We also tested the influence of the same model on the colour style of interior sketch under different colours and styles. We prepared two sets of dataset. The image style of the first dataset was bright and warm, and the image style of the second dataset was cold. Figure 10 shows the contrast effect of different styles of dataset training on the final colouring of the model.

The experimental results show that under the dataset with bright colours and sufficient interior light performance (results in Style(a)), the resulting image style tends to be warm. In a dataset with low colour saturation and low image colour temperature (results in Style(b)), the contrast of the image will decrease...
slightly, and the colour rendering will also tend to be cold. At the same time, under the premise of ensuring the diversity of style, the colouring performance and overall style of sketch did not degrade. We further confirmed that our proposed model can select different colouring effects according to different style datasets, which provides diverse style possibilities for complex interior style.

6 Conclusion

In this article, we propose VAGAN, a cycle-consistent model to transform an interior sketch photo into an interior colourised photo. We verified the impact of different datasets on the colouring results and compared the effect of each variable on the colouring results. Experiments show that using the VAM can solve the problem of insufficient datasets, improve the accuracy of colouring and the ability to capture scene style. We observe a gap between the results obtained from paired training data and unpaired training data, which may require some form of weak semantic hyper vision, or generate a more powerful translator from weak or semi-supervised data. In addition, during the test phase of the model, the colouring style of the linear artwork is often affected by the dataset. In future work, we can add multi-scale network to implement interactive colouring and colour control methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants: Methodologies for Understanding Big Data and Knowledge Discovery (61836016).

References

1. Sangkloy, P., et al.: Scribbler: Controlling deep image synthesis with sketch and color. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5400–5409 (2017)
2. Xian, W., et al.: Texturegan: Controlling deep image synthesis with texture patches. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6(1), pp. 8456–8465 (2016)
3. Gatys, L.A., et al.: Image style transfer using convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
4. Ulyanov, D., et al.: Texture networks: Feed-forward synthesis of textures and stylized images. Proceedings of the 33rd International Conference on Machine Learning, pp. 1349–1357 (2016)
5. Mirza, M., Osindero, S., Conditional generative adversarial nets. arXiv:1411.1784 (2014). https://arxiv.org/pdf/1411.1784.pdf
6. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
7. Zhu, J.-Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
8. Johnson, J., et al.: Perceptual losses for real-time style transfer and super-resolution. European Conference on Computer Vision, pp. 694–711 (2016)
9. Dumoulin, V., et al.: A learned representation for artistic style. arXiv preprint, arXiv:1610.07629, (2016). https://arxiv.org/pdf/1610.07629.pdf
10. Dosovitskiy, A., et al.: Learning to generate chairs with convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1538–1546 (2014)
11. Li, C., Wand, M.: Combining Markov random fields and convolutional neural networks for image synthesis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2479–2486 (2016)
12. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. European Conference on Computer Vision, pp. 702–716 (2016)
13. Chen, T.Q., Schmidt, M.: Fast patch-based style transfer of arbitrary style. arXiv:1612.04337, (2016). https://arxiv.org/pdf/1612.04337.pdf
14. Ashikhmin, M., Synthesizing natural textures. Proceedings of the 2001 Symposium on Interactive 3D Graphics, pp. 217–226 (2001)
15. Reinhard, E., et al.: Color transfer between images., IEEE Comput. Graph. Appl. 21(5), 34–41 (2001)
16. Welch, T., et al.: Transferring color to greyscale images. ACM Trans. Graphics 21(3), 277–280 (2003)
17. Sermanet, P., et al.: Convolutional neural networks applied to house numbers digit classification. arXiv:1204.3968, (2012). https://arxiv.org/pdf/1204.3968.pdf
18. Gregor, K., et al.: Draw: A recurrent neural network for image generation. Computer Science, pp. 1462–1471, (2015)
19. Cheng, Z., et al.: Deep colorization. Proceedings of the IEEE International Conference on Computer Vision, pp. 415–423 (2016)
20. Simonyan, K., Zisserman, A., Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, (2014). https://arxiv.org/pdf/1409.1556.pdf
21. Iizuka, S., et al.: Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Trans. Graphics. 35(4), 1–11 (2016)
22. Yao, Y., et al.: Attention-aware multi-stroke style transfer. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1467–1475 (2019)
23. Park, D.Y., Lee, K.H., Arbitrary style transfer with style-attentional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5880–5888 (2019)
24. Goodfellow, I., et al.: Generative adversarial nets. International Conference on Neural Information Processing Systems, pp. 2672–2680 (2014)
25. Radford, A., et al.: Unsupervised representation learning with deep convolutional generative adversarial networks. International Conference on Image and Graphics, pp. 97–108 (2017)
26. Hensman, P., Aizawa, K.: cGAN-based manga colorization using a single training image. 14th IAPR International Conference on Document Analysis and Recognition, pp. 72–77 (2017)
27. Furusawa, C., et al.: Comicolorization: Semi-automatic manga colorization. SIGGRAPH Asia 2017 Technical Briefs 12, pp. 1–4 (2017)
28. Zhang, L., et al.: Two-stage sketch colorization. ACM Trans. Graphs 37, 261:1–261:14 (2018)
29. Tang, H., et al.: Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2417–2426 (2019)
30. Sudhakaran, S., et al.: Lsta: Long short-term attention for egocentric action recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9954–9963 (2018)
31. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3085–3094 (2019)
32. Wang, W., et al.: Salient object detection with pyramid attention and salient edges. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1448–1457 (2019)
33. Fan, D.-P., et al.: Shifting more attention to video salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8554–8564 (2019)
34. Guo, H., et al.: Visual attention consistency under image transforms for multi-label image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 729–739 (2019)
35. Fu, J., et al.: Dual attention network for scene segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
36. Li, Y., et al.: Attention-guided unified network for panoptic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7026–7035 (2019)
37. Riaz Muhammad, U., et al.: Learning deep sketch abstraction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8014–8023 (2018)
38. Shen, Y., et al.: Zero-shot sketch-image hashing. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3598–3607 (2018)
39. Hu, C., et al.: Sketch-a-classifier: Sketch-based photo classifier generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9136–9144 (2018)
40. Dutta, A., Akata, Z.: Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5089–5098 (2019)
41. Wan, W., et al.: Transfer deep feature learning for face sketch recognition. Neural Comput. Appl. 31, 9175–9184 (2019)
42. Xu, P., et al.: Sketchmate: Deep hashing for million-scale human sketch retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8090–8098 (2018)
43. Yi, R., et al.: APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10743–10752 (2019)
44. Chen, W., Hays, J.: Sketchygan: Towards diverse and realistic sketch to image synthesis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9416–9425 (2018)
45. Ronneberger, O., et al.: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, pp. 234–241 (2015)
46. Ng, A.Y.: Feature selection, L 1 vs. L 2 regularization, and rotational invariance. Proceedings of the 21st International Conference on Machine Learning, pp. 78–79 (2014)
47. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22(10), 1345–1359 (2010)
48. Ding, L., Goshtasby, A.: On the canny edge detector. Pattern Recognit. 34(3), 721–725 (2001)
49. Ketkar, N.: Introduction to Pytorch. Apress: Berkeley, CA, (2017)
50. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980, (2014). https://arxiv.org/pdf/1412.6980.pdf
51. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 3213–3223 (2016)

How to cite this article: Li X, Li H, Wang C, Hu X, Zhang W. Visual-attention GAN for interior sketch colorisation. IET Image Process. 2021;15:997–1007. https://doi.org/10.1049/ipr2.12080