scientific reports

Check for updates

OPEN

A facial structure sampling contrastive learning method for sketch facial synthesis

Kangning Du[1,2], Jiyu Zhang[1,2], Lin Cao[1,2]✉, Yanan Guo[1,2] & Wenwen Sun[1,2]

Sketch face synthesis aims to generate sketch images from photos. Recently, contrastive learning, which maps and aligns information across diverse modalities, has found extensive application in image translation. However, when applying traditional contrastive learning to sketch face synthesis, the random sampling strategy and the imbalance between positive and negative samples result in poor performance of synthesized sketch images regarding local details. To address the above challenges, we propose A Facial Structure Sampling Contrastive Learning Method for Sketch Facial Synthesis. Firstly, we propose a region-constrained sampling module that utilizes the distribution map of facial structure obtained by a dual-branch attention mechanism to segment the input photos into diverse regions, thereby providing regional constraints for sample selection. Subsequently, we propose a dynamic sampling strategy that dynamically adjusts the sampling frequency based on the feature density in the distribution map, thereby alleviating sample imbalance. Additionally, to diminish the background influence and enhance the delineation of character contours, we introduce the mask derived from the input photo as an additional input. Finally, to further enhance the quality of the synthesized sketch images, we introduce pixel-wise loss and perceptual loss. The CUFS dataset experiment demonstrates that our method generates high-quality sketch images, outperforming existing state-of-the-art methods in subjective and objective evaluations.

Keywords Contrastive learning, Sampling strategy, Sketch facial synthesis

Sketch face synthesis aims to produce sketch images with a distinct style from provided photos, extensively utilized in both digital entertainment and criminal investigations.

In digital entertainment, sketch face synthesis methods are used to rapidly convert exaggerated face photos into sketch images with distinct styles and personalities, thereby markedly improving conversion efficiency. In criminal investigations, sketch face synthesis methods address the challenge of low identification rates between hand-drawn sketch images and images within police databases. Converting optical photos from police databases into sketched images not only enhances recognition accuracy but also improves investigation efficiency. Therefore, to improve the performance of sketch face synthesis methods in these tasks, it is crucial to develop algorithms that generate sketched face images with realistic facial details.

Present methodologies for sketch face synthesis can be broadly categorized into two groups: shallow learning-based methods and deep learning-based methods. Shallow learning-based methods can be further categorized into three primary classes: subspace learning[1], sparse representation[2], and Bayesian inference[3]. Nonetheless, these methods are limited by their inadequate generalization capability, leading to considerable variations in the generated sketch face images across diverse datasets.

In recent years, Generative Adversarial Networks (GANs) have been widely applied in sketch face synthesis. By employing adversarial strategies between the generator and discriminator, GANs notably improve the model's generalization capability, allowing for the production of high-quality sketch face images. For example, Yi et al.[4] proposed APDrawing, which synthesizes key facial features using multiple GANs. Subsequently, these synthesized facial components are stitched together to form a complete facial representation. As the local facial components are synthesized independently, inconsistencies arise in the connected regions and their appearances. To tackle this issue, Fan et al.[5] proposed FSGAN, a two-stage face synthesis framework. In the first stage, this model resembles APDrawing. In the second stage, it comprehensively incorporates global information and texture details to enhance the synthesis effect, thereby improving the quality of the synthesized sketch images. Despite its

[1]Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100101, China. [2]Key Laboratory of Information and Communication Systems, Ministry of Information Industry, Beijing Information Science and Technology University, Beijing 100101, China. ✉email: charlin26@163.com

enhanced performance, this model has also led to increased training costs. Contrastive learning has substantially enhanced training efficiency by effectively mapping and aligning information across diverse modalities, leading to remarkable advancements in the field of image translation. For example, Park et al.[6] proposed a framework based on contrastive learning. By maximizing mutual information between the source and target domain, this approach brings together corresponding samples in the source and target while pushing away noncorresponding "negative" samples, thereby enhancing image quality and reducing training time.

Traditional contrastive learning methods employ random sampling strategies that ignore the distribution of facial structure. Utilizing traditional contrastive learning methods in sketched face synthesis may produce numerous invalid samples, thus complicating model training, as shown in Fig. 1. Moreover, the imbalance between positive and negative samples leads to the loss of local texture details in the generated sketched face images. These limitations align with the findings of Bian et al.[7], who demonstrated that current sketch face recognition methods are fundamentally constrained by both generation quality deficiencies and insufficient local feature representation. To tackle these challenges, we present a contrastive learning approach based on face structure, which utilizes face structure information to precisely direct the sampling process, thereby improving the accuracy of feature extraction and opening up a new perspective for the application of contrastive learning in the field of sketch face synthesis. Specifically, we propose the Region-Constrained Sampling (RSC) module for sample selection. This module utilizes a dual-branch attention mechanism to obtain the distribution map of facial structure. Guided by this map, it segments the input photos into facial and hair regions, thereby providing regional constraints for sample selection. The RSC module effectively reduces the number of invalid samples and ensures that the selected samples contain more source domain information. Furthermore, we introduce the Dynamic Sampling Strategy (DSS), which dynamically adjusts the sampling frequency based on the feature density of samples, thus mitigating the issue of local texture loss in generated face sketch images. Additionally, we introduce a mask derived from the input photo as an extra input to the generator, reducing the influence of the background on synthesis results and enhancing the delineation of character contours. Compared with existing approaches, our proposed method demonstrates superior capability in capturing intricate facial details and structural information, enabling the synthesis of sketch faces with enhanced realism and naturalness in both texture and geometry on real-world datasets. In addition, by specifically enhancing the extraction and representation of local features, the proposed method offers a promising solution for improving the accuracy of sketch-photo cross-modal face recognition[7]. In the field of face recognition, our synthesized sketch images show a significant improvement in recognition accuracy compared to other mainstream synthesis methods. This advancement provides strong support for the practical application of sketch face synthesis technology in areas such as digital entertainment, criminal investigations.

The main contributions of this model are as follows:

- We propose A Facial Structure Sampling Contrastive Learning Method for Sketch Facial Synthesis. By imposing regional constraints during the sample selection, our method effectively addresses the issue of introducing invalid samples in traditional contrastive learning.
- We propose a Region-Constrained Sampling (RCS) module, employing a dual-branch attention mechanism to obtain the distribution map of facial structures. Guided by this map, the input photos are segmented into facial and hair regions based on feature density, thereby providing region constraints for sample selection.
- We propose a Dynamic Sampling Strategy (DSS) that dynamically adjusts the sampling quantity based on the feature density of samples, effectively addressing the imbalance between positive and negative samples.
- We innovatively introduce the mask derived from the input photo as an additional input, effectively reducing the impact of the background and enhancing the character contours of the sketch image.

Related work

Shallow learning-based methods

Subspace learning methods assume correlation between sketch face images and face photos in a low-dimensional feature space. Huang et al.[8] proposed a model to address coupled dictionary and feature space learning issues simultaneously. This model extracts feature space that not only associates cross-domain data for recognition but also updates dictionaries in each data domain to improve image representation, thereby enhancing image synthesis quality. Song et al.[9] proposed the face portrait method based on spatial portrait denoising, which adds smoothness constraints to reduce noise, effectively solving missing detail in synthetic images.

Fig. 1. Randomly sampling selected samples may include background regions unrelated to the task of synthesizing sketch images. The RCS module confines the selection region of samples to facial and hair regions, thereby minimizing irrelevant samples.

Sparse representation methods encode face images into a sparse coefficient matrix and obtain coefficients through dictionary learning. Wang et al.[10] proposed semi-coupled dictionary learning, capturing structural features of two different style images by learning a pair of dictionaries and the mapping function. This learning strategy reveals the underlying relationship between the styles, enabling precise cross-style transformation. Zhang et al.[11] proposed a method that combines the similarity between image blocks and prior knowledge. This approach employs a sparse coefficient matrix instead of traditional pixel values and extends the search area globally, effectively utilizing identity details and enhancing the realism of generated facial images.

Bayesian learning methods utilize probability models to predict outputs. Wang et al.[12] proposed the Markov Random Field (MRF) model that utilizes both local and global relationships of facial features to enhance synthesized image quality. Zhou et al.[13] proposed the Markov Weighted Field (MWF) model, which extends the one-to-one facial block transformation strategy to the one-to-many form and utilizes a weighted block method to obtain the final output blocks. This model significantly enhances facial detail realism in synthesized images, thereby improving the overall visual quality.

However, the inference-based generation of sketch images makes these algorithms time-consuming. Additionally, the quality of the generated images exhibits significant differences across different datasets.

Deep learning-based methods

Deep learning-based models handle high-dimensional data and efficiently capture rich details, significantly enhancing their generalization capability. This has opened up a new era in the field of image translation. Zhu et al.[14] proposed BiCycleGAN, which employs conditional variational autoencoder and latent regression generator to address mode collapse, thus facilitating diverse style perception. Zhang et al.[15] proposed the NPGM framework, which utilizes a probabilistic graphical model to reconstruct the common facial structure, thereby solving the problem that existing methods lose part of the facial structure during the synthesis process. Zhang et al.[16] proposed DLLRR, which transforms the sketch face synthesis task into a low-rank optimization problem, thereby synthesizing a clear and realistic sketch with identity feature information. Yu et al.[17] proposed CA-GAN, which utilizes perceptual loss function to ensure synthetic images resemble real images, and employs stacked CA-GAN further enriches synthesized images with captivating details. Zhao et al.[18] proposed ACL-GAN, which introduce adversarial consistency loss to preserve the commonalities between the source and target domains, improving image translation quality. Duan et al.[19] proposed a multi-scale gradient self-attention residual learning framework, which introduces an attention mechanism to selectively enhance key features, thereby effectively improving synthesized image quality. Liu et al.[20,21] proposed attribute-guided sketch face synthesis methods, which employ facial attribute information to generate a wider variety of local features for simulating modal differences, and embed discriminative information guided by facial attributes, thereby eliminate bias during the generation process and more effectively utilizes the attribute information. Liu et al.[22] proposed HFIDR, which learns interpretable disentangled representations via supervised disentanglement and incorporates a face semantic part exchange strategy along with symmetric adversarial loss to enhance the performance of heterogeneous face recognition and synthesis tasks. Wang et al.[23] proposed PITI, which leverages adversarial training to enhance texture synthesis in diffusion[24] models, and combines normalized guided sampling to improve synthesized image quality. Cao et al.[25] proposed a full-scale identity supervision method. This method utilizes a face recognition network to extract multi-level depth representations of cross-domain facial images and constrains the generation model using full-scale identity loss. This approach not only maximizes the preservation of perceptual appearance but also enhances the richness of synthesized images in detail. Liu et al.[26] proposed the MAMCO-HFR method that generates modality-independent perturbation samples via an adversarial training procedure, maps the data into a modality-independent subspace to mitigate the differences in data modalities and enhance the recognition performance. Bian et al.[7] proposed a novel adapter module that integrates transformer and graph convolutional network (GCN) architectures, which by enhancing the CLIP model's capability to extract features from different face regions within the same modality in sketch-photo cross-modal face recognition tasks, significantly improves cross-modal matching accuracy.

Recently, contrastive learning has been applied in image translation[6,27,28]. Its core idea is to compare positive and negative samples in the feature space to learn the feature representation of the sample. For example, Park et al.[6] proposed CUT, which utilizes contrastive learning to maximize the mutual information between input and output without relying on cycle consistency. This model outperforms existing methods in one-sided image translation tasks. Zheng et al.[27] proposed the F-LSeSim method, which introduces spatial correlation loss to effectively capture the spatial relationships within the image and eliminate the interference of visual appearance factors, thereby significantly improving the model performance. Hu et al.[28] proposed the QS-Attn method, which selectively focuses on important samples by measuring the importance of input image features, and imposes constraints that are more closely related to the translation task when calculating the contrastive loss, thereby improving the image translation quality. In addition, Oord et al.[29] proposed a method based on contrastive predictive coding. This method utilizes probabilistic contrastive loss to induce latent space, thereby better capturing the crucial information for predicting samples. He et al.[30] proposed MoCo, which facilitates contrastive unsupervised learning by constructing a dynamic dictionary with a queue and a moving-averaged encoder. Kim et al.[31] proposed InstaFormer, which leverages the self-attention mechanism in Transformer[32] to extract global content features, and enhance the quality of image conversion by defining instance-level content contrast loss between input and output images. Although we utilize patch contrastive loss in images similar to CUT, we innovatively propose a method that utilizes the distribution map of facial structure to ensure that samples can concentrate on the facial and hair regions of the input image, thereby improving the synthesis quality of sketch images.

Methods

The proposed network framework, depicted in Fig. 2, consists of a generator, discriminator, and block feature extractor, designed to produce sketch images with authentic textures from input optical images. Firstly, the optical photo and its derived mask are fed into the generator to produce the pseudo-sketch image. Subsequently, the discriminator evaluates the pseudo-sketch image to distinguish its authenticity, with feedback provided to the generator accordingly. During this process, the optical photo is fed into the pre-trained Region-Constrained Sampling (RCS) module to obtain the distribution map of facial structures. Concurrently, the generator's encoder module extracts multi-level feature maps from optical photos and pseudo-sketch images. Guided by the distribution map, the block feature extractor is employed to acquire the feature stack of the optical photos, facilitating the computation of the block contrastive loss. Notably, the introduction of the mask aims to minimize background influence and enhance character outlines in the synthesized sketch image. Finally, by integrating adversarial loss, block contrastive loss, pixel-wise loss, and perceptual loss, our model excels in producing realistic sketch images.

Dataset

The participant photos were obtained from the CUFS database[12]. It includes the Chinese University of Hong Kong (CUHK) student database[12], the AR database[33], and the XM2MTS database[34]. We have successfully secured the necessary permissions for dataset utilization. Moreover, all participants have provided their fully-informed consent to partake in this experiment. CUFS dataset can be downloaded from https://mmlab.ie.cuhk.edu.hk/datasets.html.

Region-constrained sampling module

To enhance the realism of local details in the sketch image, it is necessary to select patches containing more source domain information. Based on this, we propose the Region-Constrained Sampling (RCS) module, which employs a dual-branch attention mechanism to acquire the distribution map of the facial structures of the optical image, as shown in Fig. 3. Guided by this map, the selection region of samples is restricted to the facial and hair regions in the optical image. The RSC module effectively reduces the number of invalid samples and ensures that the selected samples contain more source domain information, aiming for maximum consistency between the synthesized sketch image and the optical image.

To emphasize the importance of each channel in the feature map, we employ the average pooling (AvgPool) layer to obtain the global feature z_c of the optical photo p_i. Subsequently, z_c is passed through the fully connected layer and activation function to compute the weight s for each channel in the feature map. The specific process is as follows:

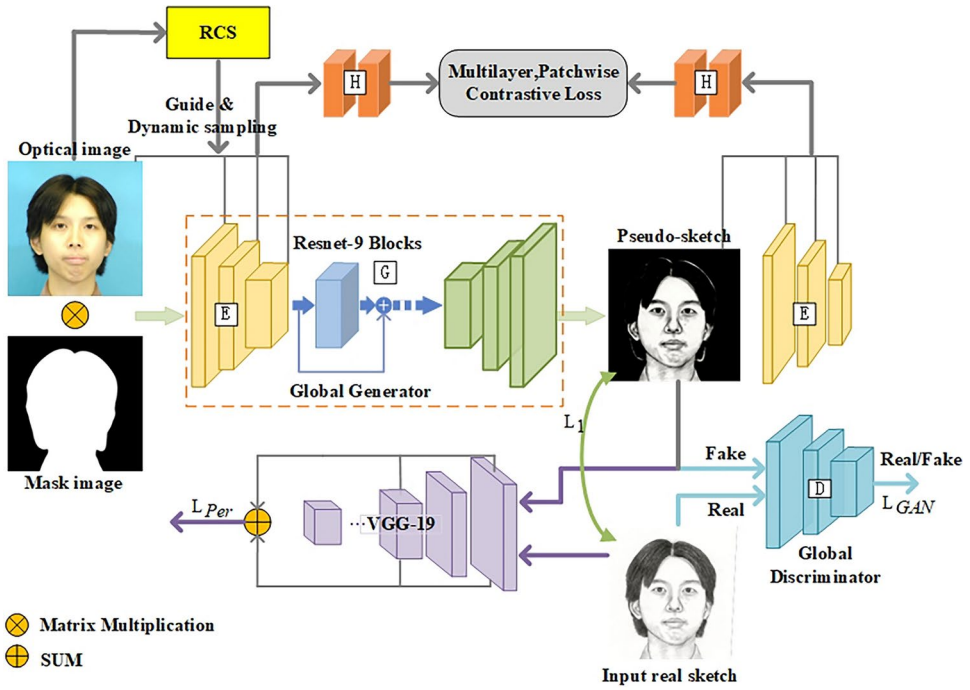

Fig. 2. The structure of A Facial Structure Sampling Contrastive Learning Method for Sketch Facial Synthesis (FSS). Here, G is the generator, E is the encoder module of the G, D is the discriminator, H is the block feature extractor, and Region-Constrained Sampling (RCS) is responsible for selecting samples with specific positional constraints as samples required for computing block contrastive loss. During the sample selection process, we introduce the Dynamic Sampling Strategy (DSS) to address the imbalance between positive and negative samples.

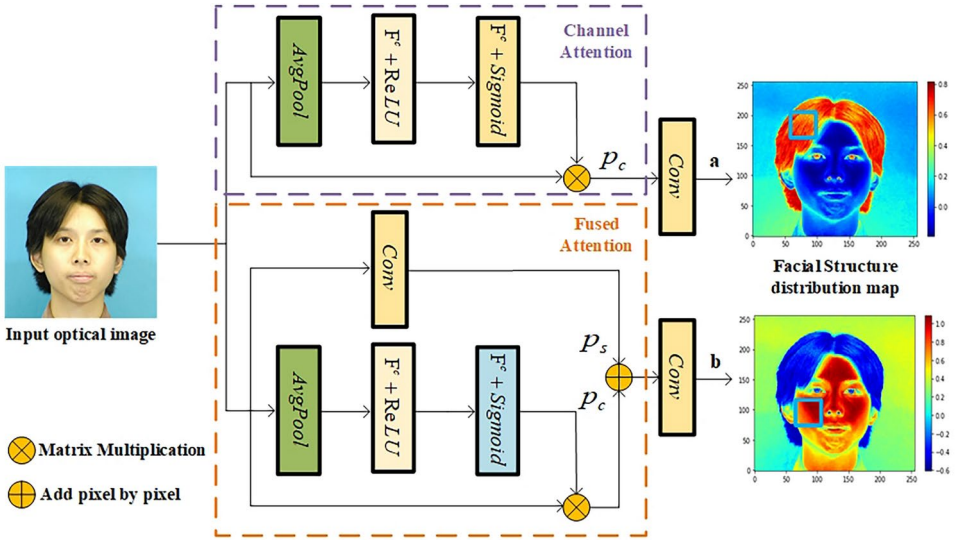

Fig. 3. The structure of the region-constrained sampling (RCS) module. A and b, respectively, represent the attention maps of optical photos obtained under different branches. The depth of red indicates the degree of attention the branch pays to the region, with deeper red indicating higher attention to that region.

$$z_c = AvgPool(p_i) \tag{1}$$

$$s = Sigmoid(F^c(ReLU(F^c(z_c)))) \tag{2}$$

Where F^c represents the fully connected layer, *ReLU* and *Sigmoid* represent the corresponding activation functions.

To improve the representation of the hair region in the optical photo, we employ the vector s to excite the optical photo p_i. Then the resulting output undergoes a 1×1 convolution layer to obtain the enhanced channel attention matrix $(p_i)_c$, as shown in Eq. (3).

$$(p_i)_c = Conv(sp_i) \tag{3}$$

To assess the performance of channel attention in the optical photos, we visualize the attention matrix $(p_i)_c$, as shown in Fig. 3a. Through visualization analysis, distinct performances across different regions can be observed under the channel attention mechanism. In the hair region, the color and texture of the hair create prominent feature patterns within specific channels. The channel attention mechanism is capable of sensitively detecting these patterns and assigning higher weights to the corresponding channels, which effectively accentuates the hair region. Conversely, in the facial region, the facial features and their relative positional relationships play a crucial role. Nevertheless, the channel attention mechanism fails to take into account the spatial relationships. As a result, it is unable to effectively capture this vital information, leading to suboptimal performance in the facial region.

To improve the spatial information of the feature map, we first subject the original feature map p_i to a 1×1 convolution to produce the compression matrix z_s. Subsequently, we normalize the compression matrix z_s using the *Sigmoid* activation function and then multiply it with the original feature map p_i to obtain the enhanced spatial feature matrix $(p_i)_s$. The specific process is as follows:

$$z_s = Conv(p_i) \tag{4}$$

$$(p_i)_s = Sigmoid(z_s)p_i \tag{5}$$

The channel attention implementation is shown in Eqs. (1, 2, and 3). Finally, we feed the enhanced spatial and channel feature matrix into a 1×1 convolutional layer to generate the fused attention matrix f_f, as shown in Eq. (6):

$$f_f = Conv((p_i)_s + (p_i)_c) \tag{6}$$

To assess the performance of fused attention, we visualize the attention matrix f_f, as shown in Fig. 3b. The fused attention mechanism integrates multiple attention modes and comprehensively analyzes the facial region from two dimensions: spatial and channel. This enables the mechanism to accurately capture the overall characteristics of the face and the relative positional relationships between facial features. Consequently, it allocates more attention to the facial region, achieving effective representation of facial features. In contrast, characteristics such as color and texture in the hair region are predominantly captured by the channel attention mechanism.

The spatial attention mechanism often fails to effectively enhance such channel-based features. Furthermore, the introduction of the spatial attention mechanism into the fused attention mechanism alters the overall attention distribution pattern. As a result, under the fused attention mechanism, the performance in the hair region may not achieve the ideal effectiveness observed when using the channel attention mechanism alone.

In summary, the RCS module employs distinct branches to obtain attention maps of the optical photo, effectively segmenting it into facial and hair regions, thereby offering regional constraints for sample selection.

Dynamic sampling strategy

In the sketch face synthesis task, we confine the selection region of contrastive learning samples to the hair and facial regions, effectively mitigating the interference of background information on the synthesized results. This strategic not only refines the learning direction of the model but also substantially enhances the model's efficiency in extracting key features during training. Consequently, it reduces the training complexity, rendering the overall process more precise and efficient.

During the sample selection procedure, we initially leverage the encoder module of the generator to extract multi-level feature maps from the optical photo. Subsequently, we utilize the facial structure distribution map generated by the RCS module as a regional constraint to conduct sample selection within the extracted multi-level feature maps. Notably, we discovered that the multi-level feature maps extracted by the encoder of the generator are capable of effectively retaining the key identity information of the characters. Consequently, we carry out the sample selection operation on the multi-level feature maps employing the geometric scaling method. Next, following the methodology of SimCLR[35], we process the acquired samples through the block feature extractor H to obtain the feature stack of the optical photo. The feature stack of the synthetic sketch image can be acquired using an same approach. Moreover, in the sample selection process, we introduce a Dynamic Sampling Strategy (DSS), this strategy dynamically modulates the sampling frequency in accordance with the feature density of samples in the distribution map, alleviating the imbalance between positive and negative samples and preventing the distortion of local details in the synthetic sketch image. We utilize the obtained feature stack to compute the multi-layer block contrast loss, as shown in Fig. 4. The specific implementation steps are as follows:

To produce the feature stack of the optical image p, we utilize the encoder module of the generator to extract multi-level feature maps from the optical photo p. The extracted feature maps are fed into the block feature extractor H guided by the RCS module, thereby producing a feature stack $\{z_l\}_L$ with positional constraints. The specific process is shown in Eq. (7).

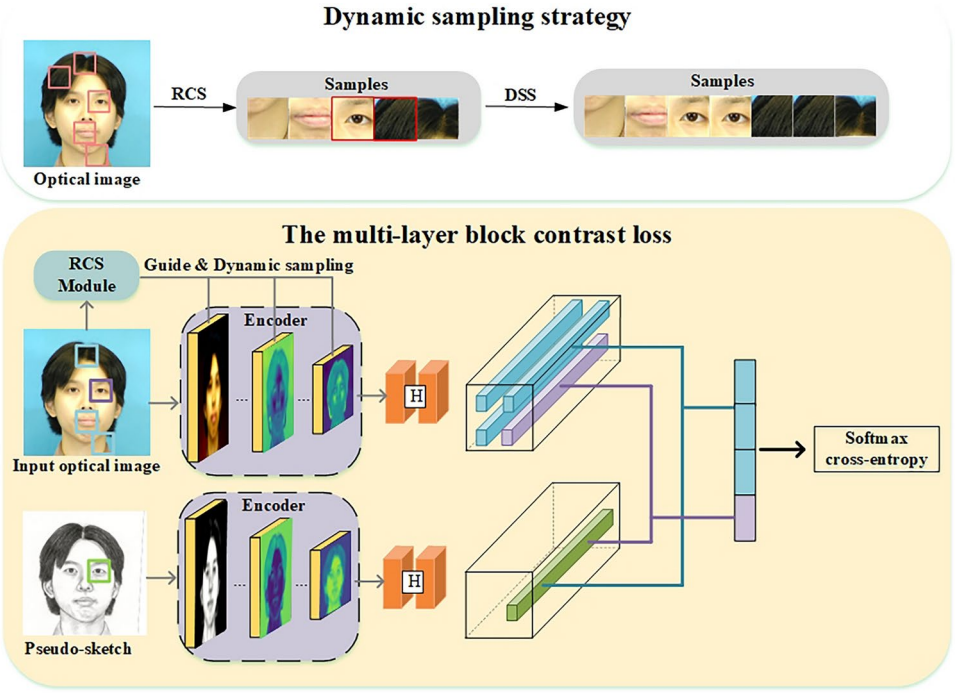

Fig. 4. The dynamic sampling strategy demonstrates the mechanism for adjusting the number of samples. Under the constraint of RSC, the samples are dynamically adjusted based on their feature density in the face structure distribution map. Specifically, the orange boxes represent the obtained samples, while the red boxes indicate samples whose feature density exceeds the preset threshold, thereby the number of such samples being doubled. The calculation procedure of the multi-layer block contrast loss demonstrates how the multi-layer feature maps of optical photos and sketch images are converted into corresponding feature stacks under the guidance of the RSC module. In the figure, the query sample is denoted by the green box, the positive sample by the purple box, and the negative samples by the blue box.

$$\{z_l\}_L = H(RCS(E^l(p))) \tag{7}$$

Here, E^l denotes the output feature of the encoder at layer l where $l \in \{0, 4, 8, 12, 16\}$.

Similarly, the encoder extracts feature stacks $\{\hat{z}_l\}_L$ with region constraints from pseudo-sketch images $\hat{s}$, as shown in Eq. (8):

$$\{\hat{z}_l\}_L = H(RCS(E^l(\hat{s}))) \tag{8}$$

We sequentially traverse each feature in feature stacks $\{\hat{z}_l\}_L$ and $\{z_l\}_L$. During this process, we label different features: features in stack $\{\hat{z}_l\}_L$ are denoted as query samples v, those in stack $\{z_l\}_L$ as positive samples v^+, and the remaining ones in stack $\{z_l\}_L$ as negative samples v_n^-. It is worth noting that the higher the feature density of samples, the more corresponding samples there are. The positive samples we constructed have features that highly match those of the query samples, thus providing positive learning references for the model. In contrast, the features of negative samples are quite different from those of the query samples, and this difference enables them to assist the model in effectively distinguishing different features, thereby enhancing the model's discrimination ability.

To maximize the correlation between the query sample and the positive sample, we introduce the cross-entropy loss function, as shown in Eq. (9):

$$\mathrm{l}(v, v^+, v^-) = -\log\left[\frac{\exp(v \cdot v^+/\tau)}{\exp(v \cdot v^+/\tau) + \sum_{n=1}^{N} \exp(v \cdot v_n^-/\tau)}\right] \tag{9}$$

To compute the correspondence between different patches in feature maps at various scales. We introduce the multi-level patchwise contrastive loss, as shown in Eq. (10):

$$\mathrm{L_{patch}}(\mathrm{E}^l, H, \mathrm{P}) = \mathrm{E}_{p_i \sim P_P} \sum_{l=1}^{L} \sum_{s=1}^{S_l} l(\widehat{\mathrm{z}}_l^s, z_l^s, z_l^{\frac{S}{s}}) \tag{10}$$

Where, $\widehat{z}_l^s$ denotes the pseudo-sketch query sample; z_l^s denotes the optical photo positive sample; $z_l^{\frac{S}{s}}$ denotes the optical photo negative sample; $s = 1, 2, 3, \ldots, S_l$, and S_l represents the number of patches at each layer.

Loss function

Given a pair of photo-sketch sets $\{(p_i, s_i)|p_i \in P, s_i \in S\}$, with P and S denoting collections of facial photos and sketch images respectively, and the shared label i indicating that the facial photo and sketch image belong to the same individual. To ensure stability and accelerate convergence during network training, we designed the following loss functions:

Adversarial loss. During the training process of the GAN, the discriminator D aims to accurately distinguish between real sketch images and pseudo-sketch images. Meanwhile, the generator G generates pseudo-sketch images closely resembling the real sketch images to deceive the discriminator D. To quantify this adversarial process, we introduce the adversarial loss function[36] as shown in Eq. (11):

$$L_{\mathrm{GAN}}(G, D, P, S) = \mathbb{E}_{s_i \sim P_S}[\log(D(s_i))] + \mathbb{E}_{p_i \sim P_P}[\log(1 - D(G(p_i)))] \tag{11}$$

Where P_S and P_P denote the sample distributions of sketch images and optical photos, respectively.

Pixel-wise loss & Perceptual loss. The training dataset comprises paired photo-sketch images, so we employ supervised learning to train the GAN. Hence, we introduce the following loss functions: pixel-wise loss and perceptual loss.

To guide the model in capturing image details and generating high-quality sketch images, we introduce the pixel-wise loss, calculated by comparing the L1 distance between the pseudo-sketch image and the real sketch image at each pixel, as shown in Eq. (12):

$$L_1(G_s, P, S) = \mathbb{E}_{s_i \sim P_S, p_i \sim P_P}[\|G_s(p_i) - s_i\|_1] \tag{12}$$

To ensure that the model considers the global structure of the image and produces natural sketch images, we introduce perceptual loss. We utilize pre-trained VGG19[37] to extract multi-layer feature maps from both real and pseudo-sketch, computing the L1 norm of these feature maps at different layers, as shown in Eq. (13):

$$L_{\mathrm{per}}(G_s, P, S) = E_{s_i \sim P_S, p_i \sim P_P} \sum_{j=0}^{4} \omega_j \left\|\varphi_j(s_i) - \varphi_j(G_s(p_i))\right\|_1 \tag{13}$$

Where, $\varphi_j(.)$ denotes the perceptual features extracted from the $j-th$ layer, with $j = 0, 1, 2, 3, 4$. Additionally, $\omega_j = 1/32, 1/16, 1/8, 1/4, 1$.

Patchwise contrastive loss. To maximize the mutual information between optical photos and synthetic sketch images, we introduce the multi-level patchwise contrastive loss, as shown in Eq. (10).

Dataset	Train	Test
CUHK	88	100
AR	80	43
XM	100	195

Table 1. Train set and test set division of the dataset.

Number	Baseline	W/ RCS	W/ DSS	W/ Mask	W/ losses
1	✓				
2	✓	✓			
3	✓	✓	✓		
4	✓	✓	✓	✓	
5	✓	✓	✓	✓	✓

Table 2. Ablation studies.

Total loss.The total loss function is given by Eq. (14):

$$L_{full} = L_{adv} + \alpha L_1 + \beta L_{patch} + \gamma L_{per} \tag{14}$$

Where α, β, and γ are hyperparameters controlling the importance of per-pixel loss, patchwise contrastive loss, and perceptual loss, respectively.

Experimental results and analysis

This section begins by detailing the experimental datasets, evaluation metrics, and implementation specifics. We then validate the effectiveness and accuracy of our proposed method using the publicly available CUFS dataset, conducting both qualitative and quantitative experiments.

Training and testing dataset: We utilize the public CUFS dataset, comprising 188 pairs of faces from the Chinese University of Hong Kong (CUHK) student dataset, 123 pairs from the AR dataset, and 295 pairs from the XM2VTS dataset. Both the paired photos and sketch images feature neutral facial expressions. They are resized to 256 $\times$ 256 and geometrically aligned based on the eyes. To facilitate comparison with existing methods, we follow Wang et al.[38] dataset division method. Specifically, 88 pairs from CUHK, 80 pairs from AR, and 100 pairs from XM2VTS are used for training, while the remaining pairs are reserved for testing. Data division details for each dataset are shown in Table 1.

Evaluation metrics: To evaluate the performance of the proposed method, we employ Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM)[39], and Fréchet Inception Distance (FID)[40]. A higher PSNR and SSIM value, along with a lower FID value, collectively indicate the superior quality of the pseudo-sketch image.

Implementation details: The proposed method is implemented using the PyTorch framework. During training, the batch_size is set to 1, and the total number of iterations is 400 epochs. The learning rate starts at 0.0002 for the first 200 epochs and linearly decays to 0 for the remaining 200 epochs. The generator, discriminator, and block feature extractor are optimized using the Adam optimizer, with momentum parameters β_1 and β_2 set to 0.5 and 0.999, respectively. The weights α and β for the per-pixel loss and perceptual loss in the total loss function are set to 100 and 1, respectively. Furthermore, the number of samples for the face and hair regions, guided by the facial structure distribution, are set to 256 and 128, respectively.

Analysis of ablation experimental results

To verify the effectiveness of each component in the proposed method, a series of ablation experiments are conducted. Starting with the original CUT model as the baseline, we incrementally introduced the Region-Constrained Sampling (RCS) module, Dynamic Sampling Strategy (DSS), mask, pixel-wise and perceptual losses during training. The performance of these experiments on the CUFS dataset is presented in Table 2.

As shown in Fig. 5, CUT employs random sampling to synthesize sketch images, which falls short in capturing facial features like eyes and textures on the CUHK and AR datasets. Similarly, on the XM2VTS dataset, despite the depiction of character outlines and facial features, the synthesized sketch images exhibit dim colors, failing to convey facial texture and shadow distribution, and thus lacking realism. To address these challenges, we introduced the RCS module. On the AR and XM2VTS datasets, the synthesized sketch images show significant improvements over CUT in terms of hair texture, facial details, and line strokes. On the CUHK dataset, while the details of the eyes and mouth are enhanced, issues such as artifacts in character outlines and missing key lines remain pressing problems. Additionally, across all datasets, the synthesized sketch images exhibit local discrepancies compared to real sketch images. To further enhance the model's performance in capturing local details, we introduced the DSS. This strategy effectively mitigates artifacts and missing lines on the CUHK dataset, making the synthesized sketch images more realistic in terms of eye and hair textures. However, the newly added lines fail to align with human visual aesthetics. On the AR and XM2VTS datasets, the

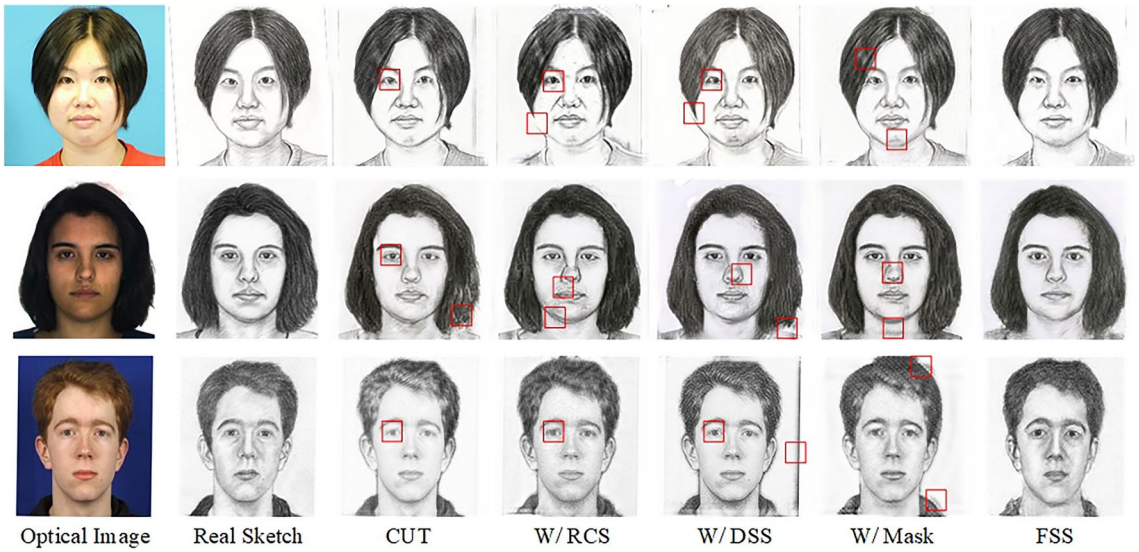

Fig. 5. Comparison of synthetic effects of ablation experiments.

Dataset	Metrics	CUT	W/ RCS	W/ DSS	W/ Mask	FSS
CUHK	FID↓	81.2447	77.4903	71.34153	62.7373	**53.3347**
	PSNR↑	16.8903	17.1985	17.4079	17.4951	**18.5729**
	SSIM↑	0.54361	0.57214	0.58010	0.58509	**0.63567**
AR	FID↓	87.6897	84.6346	81.2749	71.6501	**62.33**
	PSNR↑	17.2912	17.8219	17.9628	17.8549	**18.1293**
	SSIM↑	0.59046	0.60303	0.59908	0.60337	**0.63419**
XM2VTS	FID↓	76.0358	73.4575	71.5189	60.6443	**30.4706**
	PSNR↑	15.9739	16.4856	16.9673	17.1576	**18.1112**
	SSIM↑	0.50383	0.53392	0.53813	0.55885	**0.57660**

Table 3. Comparison of quantitative metrics of ablation experiments.

synthesized sketch images demonstrate improved of eyes and nose details but still suffer from localized feature loss and interference from background lines. To eliminate background interference and enhance the generator's expressiveness in character areas, we incorporated a Mask corresponding to the optical photo as an additional input to the generator. This approach results in sketch images with clearer hair textures and more realistic facial details across all datasets, though minor imperfections persist in specific areas. Furthermore, to enhance visual fidelity, we introduced pixel-wise loss and perceptual loss functions. These loss functions not only account for pixel-level differences between synthesized and real sketch images but also guide the synthesized sketch images to achieve greater consistency with real sketch images in terms of overall structure and contextual coherence.

As shown in Table 3, with the best-performing results highlighted in bold. The gradual introduction of each component in experiments, followed by comparison with CUT, it can be observed that the FID is decreased by approximately 25 on the CUHK and AR datasets and by approximately 45 on the XM2VTS dataset. Furthermore, improvements in SSIM and PSNR are evident across all datasets. These results highlight the critical role of these components in enhancing the overall performance of the proposed method.

In summary, comparing individual ablation experiments reveals that the FSS consistently produces higher-quality sketch images.

Analysis of comparative experimental results

To verify the effectiveness of our proposed method, we compare it with mainstream methods on the CUFS datasets.

As shown in Figs. 6, 7, 8. CycleGAN optimizes the synthesis of sketch images via cycle-consistency loss, which aids in capturing better global information. However, it tends to produce unrealistic textures in local regions, such as the eyes, nose, and mouth, resulting in suboptimal visual outcomes. Pix2pix relies on L1 loss for optimization, but due to the lack of precise alignment between optical photos and sketch images, the produced sketch images often exhibit facial distortions with blurred character contours. MDAL employs multi-domain adversarial learning, effectively mitigating issues related to blurriness and deformation in synthesized sketch images. However, it falls short of enhancing texture details, resulting in excessively smooth textures and

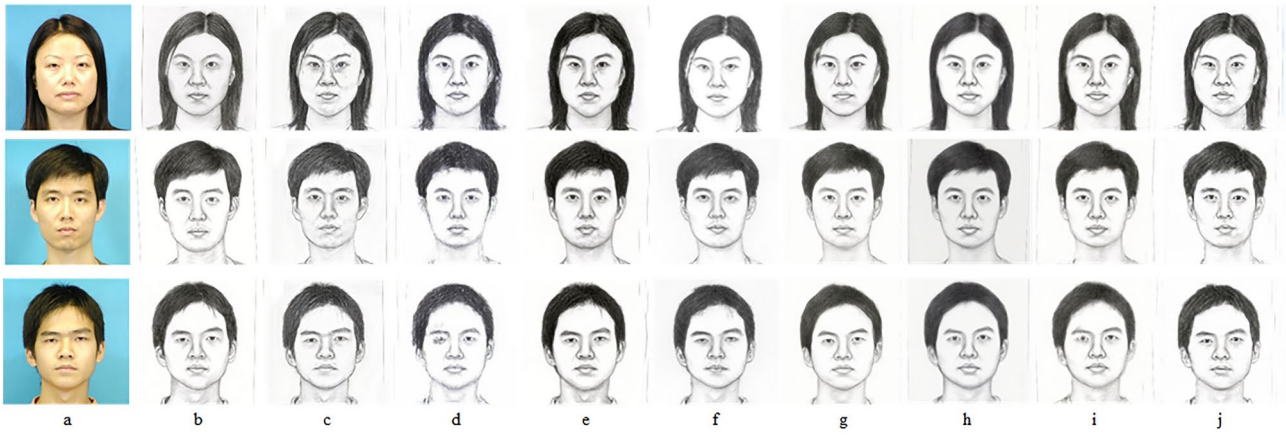

Fig. 6. Visual comparison on the CUHK test set. (**a**) Images. (**b**) Ground true. (**c**) CycleGAN[41], (**d**) pix2pix[42], (**e**) MDAL[43], (**f**) FaceSketchWild[44], (**g**) FSGAN[5], (**h**) T2V[45], (**i**) Dif-Fusion[46] and (**j**) our FSS.

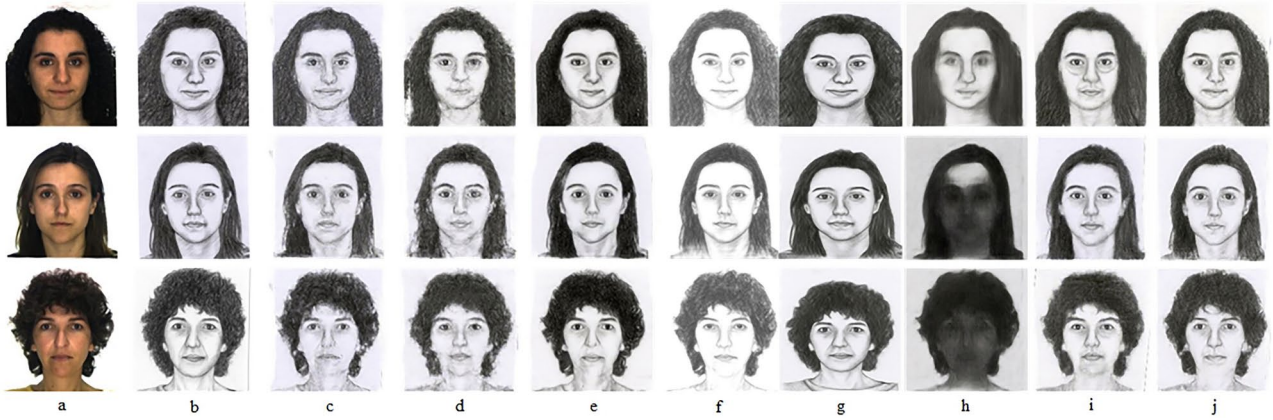

Fig. 7. Visual comparison on the AR test set. (**a**) Images. (**b**) Ground true. (**c**) CycleGAN[41], (**d**) pix2pix[42], (**e**) MDAL[43], (**f**) FaceSketchWild[44], (**g**) FSGAN[5] (**h**) T2V[45], (**i**) Dif-Fusion[46] and (**j**) our FSS.

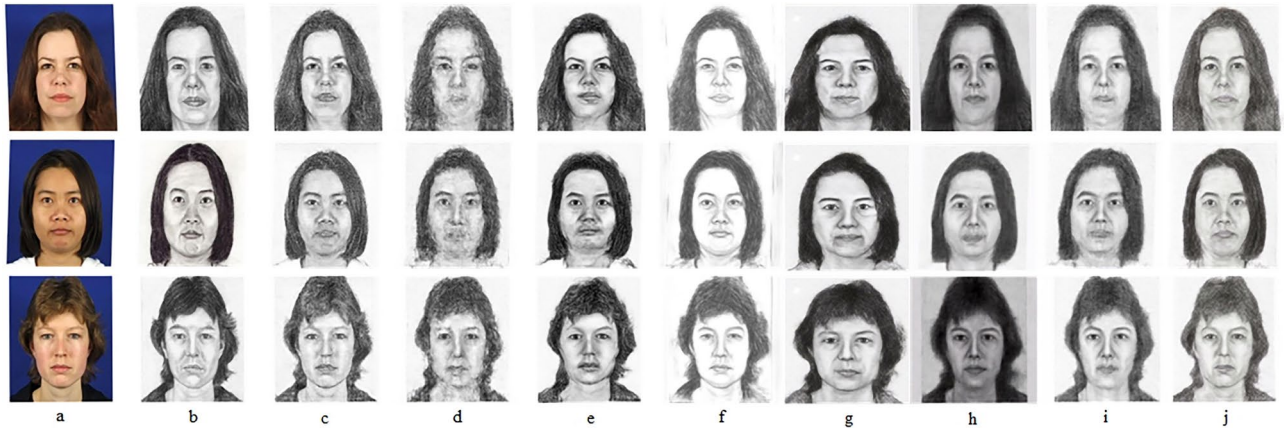

Fig. 8. Visual comparison on the XM2VTS test set. (**a**) Images. (**b**) Ground true. (**c**) CycleGAN[41], (**d**) pix2pix[42], (**e**) MDAL[43], (**f**) FaceSketchWild[44], (**g**) FSGAN[5] (**h**) T2V[45], (**i**) Dif-Fusion[46] and (**j**) our FSS.

Metrics	CycleGAN[41]	pix2pix[42]	MDAL[43]	FaceSketchWild[44]	FSGAN[5]	T2V[45]	Dif-Fusion[46]	FSS
FID↓	60.13357	74.96257	66.87124	79.86677	65.6648	89.4926	72.2763	**53.3347**
SSIM↑	0.589582	0.539621	0.635232	0.631354	**0.65022**	0.6423	0.6021	0.63567
PSNR↑	17.88922	17.68544	17.90861	15.58399	18.6751	17.4948	**18.8142**	18.5729

Table 4. Quantitative comparison on the CUHK test set.

Metrics	CycleGAN[41]	pix2pix[42]	MDAL[43]	FaceSketchWild[44]	FSGAN[5]	T2V[45]	Dif-Fusion[46]	FSS
FID↓	74.42147	103.6566	81.19116	72.41438	64.9136	110.6717	82.1047	**62.33**
SSIM↑	0.60301	0.572184	0.626738	0.61378	**0.64317**	0.5568	0.5412	0.634019
PSNR↑	18.04683	17.57163	18.04985	**19.4711**	18.46683	18.1123	18.1656	18.12983

Table 5. Quantitative comparison on the AR test set.

Metrics	CycleGAN[41]	pix2pix[42]	MDAL[43]	FaceSketchWild[44]	FSGAN[5]	T2V[45]	Dif-Fusion[46]	FSS
FID↓	38.50704	70.60789	44.41342	56.84637	44.4891	62.8700	55.0305	**30.4706**
SSIM↑	0.538109	0.517199	0.54752	0.476643	0.47257	0.5196	0.5266	**0.5766**
PSNR↑	17.37244	17.73103	16.15575	15.93193	16.4345	15.0301	18.0176	**18.1112**

Table 6. Quantitative comparison on the XM2VTS test set.

ambiguous details in the synthesized sketch images. FaceSketchWild employs a cascading approach to delineate characters' contours and key facial features. However, the color of synthesized sketch images is dim, thereby failing to effectively depict facial texture and shadow distribution, lacking realism. FSGAN adopts a two-stage facial synthesis framework, which fully considers key facial features, global information, and texture details, thereby notably enhancing the synthesis quality of sketch images. However, on the CUHK dataset, the clarity of texture in the hair region remains insufficient. Furthermore, notable disparities are observed between the synthesized sketch images and the original sketch images on the XM2VTS and AR datasets. T2V employs a conditional denoising probability model and achieves relatively satisfactory results on the CUHK and XM2VTS datasets. However, it still faces issues with suboptimal denoising performance across datasets, leading to black backgrounds or patch artifacts in some synthesized sketch images, particularly pronounced in the AR dataset. Dif-Fusion construct a multi-channel data distribution using a diffusion model, effectively mitigating the black background or patch artifacts observed in T2V. Nevertheless, the synthesized sketch images still exhibit shortcomings, such as blurred local details and dim brush strokes. In contrast to previous methods, our proposed method achieves a harmonious equilibrium among global features, local features, character contours, and visual effects within synthesized sketch images, thereby rendering the produced sketch images more realistic and artistic.

As shown in Table 4, 5, and 6, with the best-performing results highlighted in bold. It can be observed that our method surpasses CycleGAN, pix2pix, MDAL and T2V across all metrics on diverse datasets. Noteworthy is its outstanding performance on the XM2VTS dataset, where it consistently outperforms all other methods across all metrics. However, on the CUHK and AR datasets, although our method slightly trails behind FSGAN, FaceSketchWild and Dif-Fusion in terms of PSNR or SSIM, our method exhibits superior visual quality compared to these three methods.

In summary, our approach has demonstrated significant advancements in addressing issues of local blurriness in synthesized sketch images, alongside enhancing the delineation of character contours. Moreover, the method adeptly preserves subtle texture details and overall facial structure within sketch images, thereby augmenting the visual realism of the synthesized sketch images.

Face sketch recognition

Face recognition performance is a key metric for evaluating sketch face synthesis methods. To assess the performance of our proposed method in the face recognition task, we selected the FaceNet[47] model, which is based on triplet loss optimization, and conducted experimental evaluation using the CUFS dataset. FaceNet minimizes the triplet loss function to cluster embedding vectors of the same identity while effectively distinguishing embeddings of different identities. As shown in Table 7, with the best-performing results highlighted in bold. It can be observed that our method significantly outperforms other models in terms of recognition rate, highlighting its ability to better preserve identity information during the synthesis process.

Conclusions

We propose a contrastive learning approach based on face structure for sketch face synthesis. Firstly, we design the RCS module to obtain distribution maps of facial structures from optical photos. Using this map as guidance,

Comparison methods	Test set	CycleGAN[41]	pix2pix[42]	MDAL[43]	T2V[45]	Dif-Fusion[46]	FSS
Accurary (%)	85.20	88.86	88.82	90.22	68.48	89.56	**90.54**

Table 7. Comparison of recognition accuracy for synthesizing sketch images on CUFS database(%).

we segment the input photos into facial and hair regions to provide regional constraints, addressing the issue of invalid samples introduced by random sampling in traditional contrastive learning. Additionally, we propose a dynamic sampling strategy that effectively mitigates the imbalance between positive and negative samples by dynamically adjusting the sample quantity based on the feature density. Furthermore, to eliminate background influence and enhance the character contours of the sketch image, we introduce the mask derived from the optical photo as additional input. Finally, we adopt pixel-wise loss and perceptual loss, which not only capture pixel-level disparities but also consider the global structure and contextual information between the synthetic sketch image and the optical sketch image. Extensive experimentation on the CUFS dataset demonstrates our method's superior performance compared to mainstream methods, offering valuable insights and benchmarks for future investigations within sketch face synthesis domain. Looking ahead, we plan to explore the application of hard example mining in contrastive learning to further enhance the quality of sketch face synthesis.

Data availability

The datasets synthesized during and/or analysed during the current study are available from the corresponding author on reasonable request. The code can be found under DOI 10.5281/zenodo.14724243.

Received: 24 April 2024; Accepted: 29 April 2025
Published online: 08 May 2025

References

1. Huang, D. A. & Frank Wang, Y. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In *IEEE* (2013).
2. Tang, X. & Wang, X. Face photo recognition using sketch. In: *International Conference on Image Processing* (2002).
3. Chen, H., Xu, Y. Q., Shum, H. Y., Zhu, S. C. & Zheng, N. N. Example-based facial sketch generation with non-parametric sampling. In: *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on* (2001).
4. Yi, R., Liu, Y.-J., Lai, Y.-K. & Rosin, P. L. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10743–10752 (2019).
5. Fan, D.-P. et al. Facial-sketch synthesis: A new challenge. *Machine Intell. Res.* **19**, 257–287 (2022).
6. Park, T., Efros, A. A., Zhang, R. & Zhu, J.-Y. Contrastive learning for unpaired image-to-image translation. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 319–345 (Springer, 2020).
7. Bian, H., Lv, B., Guo, Y., Zhang, B. & Du, K. Sketch face recognition method based on local-global adapter. *IEEE Access* (2025).
8. Huang, D.-A. & Wang, Y.-C. F. Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition. In: *Proceedings of the IEEE international conference on computer vision*, 2496–2503 (2013).
9. Song, Y., Bao, L., Yang, Q. & Yang, M.-H. Real-time exemplar-based face sketch synthesis. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 800–813 (Springer, 2014).
10. Wang, S., Zhang, L., Liang, Y. & Pan, Q. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: *2012 IEEE Conference on computer vision and pattern recognition*, 2216–2223 (IEEE, 2012).
11. Zhang, S., Gao, X., Wang, N., Li, J. & Zhang, M. Face sketch synthesis via sparse representation-based greedy search. *IEEE Trans. Image Process.* **24**, 2466–2477 (2015).
12. Wang, X. & Tang, X. Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 1955–1967 (2008).
13. Zhou, H., Kuang, Z. & Wong, K.-Y. K. Markov weight fields for face sketch synthesis. In: *2012 IEEE conference on computer vision and pattern recognition*, 1091–1097 (IEEE, 2012).
14. Zhu, J.-Y. et al. Toward multimodal image-to-image translation. *Adv. Neural Inform. Process. Syst.***30** (2017).
15. Zhang, M., Wang, N., Li, Y. & Gao, X. Neural probabilistic graphical model for face sketch synthesis. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 2623–2637 (2019).
16. Zhang, M., Wang, N., Li, Y. & Gao, X. Deep latent low-rank representation for face sketch synthesis. *IEEE Trans. Neural Netw. Learn. Syst.* **30**, 3109–3123 (2019).
17. Yu, J. et al. Toward realistic face photo-sketch synthesis via composition-aided gans. *IEEE Trans. Cybern.* **51**, 4350–4362 (2020).
18. Zhao, Y., Wu, R. & Dong, H. Unpaired image-to-image translation using adversarial consistency loss. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 800–815 (Springer, 2020).
19. Duan, S., Chen, Z., Wu, Q. J., Cai, L. & Lu, D. Multi-scale gradients self-attention residual learning for face photo-sketch transformation. *IEEE Trans. Inf. Forensics Secur.* **16**, 1218–1230 (2020).
20. Liu, D., Gao, X., Wang, N., Li, J. & Peng, C. Coupled attribute learning for heterogeneous face recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 4699–4712 (2020).
21. Liu, D., Gao, X., Wang, N., Peng, C. & Li, J. Iterative local re-ranking with attribute guided synthesis for face sketch recognition. *Pattern Recogn.* **109**, 107579 (2021).
22. Liu, D., Gao, X., Peng, C., Wang, N. & Li, J. Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 5611–5625 (2021).
23. Wang, T. et al. Pretraining is all you need for image-to-image translation. arXiv preprint arXiv:2205.12952 (2022).
24. Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural. Inf. Process. Syst.* **34**, 8780–8794 (2021).
25. Cao, B., Wang, N., Li, J., Hu, Q. & Gao, X. Face photo-sketch synthesis via full-scale identity supervision. *Pattern Recogn.* **124**, 108446 (2022).
26. Liu, D. et al. Modality-agnostic augmented multi-collaboration representation for semi-supervised heterogenous face recognition. In: *Proceedings of the 31st ACM International Conference on Multimedia*, 4647–4656 (2023).
27. Zheng, C., Cham, T.-J. & Cai, J. The spatially-correlative loss for various image translation tasks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16407–16417 (2021).
28. Hu, X. et al. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18291–18300 (2022).

29. Oord, A. V. D., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018).
30. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738 (2020).
31. Kim, S., Baek, J., Park, J., Kim, G. & Kim, S. Instaformer: Instance-aware image-to-image translation with transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18321–18331 (2022).
32. Han, K. et al. Transformer in transformer. *Adv. Neural. Inf. Process. Syst.* **34**, 15908–15919 (2021).
33. Martinez, A. & Benavente, R. *The ar face database: Cvc technical report, 24* (Tech, Rep, 1998).
34. Messer, K. et al. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, vol. 964, 965–966 (Citeseer, 1999).
35. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*, 1597–1607 (PMLR, 2020).
36. Goodfellow, I. et al. Generative adversarial nets. *Adv. Neural Inform. Process. Syst.***27** (2014).
37. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
38. Wang, N., Gao, X. & Li, J. Random sampling for fast face sketch synthesis. *Pattern Recogn.* **76**, 215–227 (2018).
39. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
40. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inform. Process. Syst.***30** (2017).
41. Zhu, J. Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE* (2017).
42. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134 (2017).
43. Zhang, S., Ji, R., Hu, J., Lu, X. & Li, X. Face sketch synthesis by multidomain adversarial learning. *IEEE Trans. Neural Netw. Learn. Syst.* **30**, 1419–1428 (2018).
44. Chen, C., Liu, W., Tan, X. & Wong, K.-Y. K. Semi-supervised learning for face sketch synthesis in the wild. In: *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14*, 216–231 (Springer, 2019).
45. Nair, N. G. & Patel, V. M. T2v-ddpm: Thermal to visible face translation using denoising diffusion probabilistic models. In: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–7 (IEEE, 2023).
46. Yue, J., Fang, L., Xia, S., Deng, Y. & Ma, J. Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing* (2023).
47. Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823 (2015).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62201066, U20A20163).

Author contributions

Kangning Du: contributed to conceptualization, methodology, writing-review & editing. Jiyu Zhang: contributed to methodology, software, writing-original draft, investigation. Cao Lin: contributed to formal analysis, investigation, Data curation. Yanan Guo: contributed to writing-review & editing, supervision. Wenwen Sun: contributed to writing-review & editing, supervision.

Declarations

Competing interest

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00574-6.

Correspondence and requests for materials should be addressed to L.C.