Stylized Image Generation based on Music-image Synesthesia Emotional Style Transfer using CNN Network

Baixi Xing[1*], Jian Dou[2], Qing Huang[2], and Huahao Si[2]
[1] Zhejiang University of Technology
Hangzhou, China
[e-mail: xingbaixi@zjut.edu.cn]
[2] Hangzhou Dianzi University
Hangzhou, China
[e-mail: 1186400034@qq.com, huangqing@hdu.edu.cn, 2274749769@qq.com]
*Corresponding author: Baixi Xing

Received October 22, 2020; revised February 17, 2021; March 24, 2021; published April 30, 2021

Abstract

Emotional style of multimedia art works are abstract content information. This study aims to explore emotional style transfer method and find the possible way of matching music with appropriate images in respect to emotional style. DCNNs (Deep Convolutional Neural Networks) can capture style and provide emotional style transfer iterative solution for affective image generation. Here, we learn the image emotion features via DCNNs and map the affective style on the other images. We set image emotion feature as the style target in this style transfer problem, and held experiments to handle affective image generation of eight emotion categories, including dignified, dreaming, sad, vigorous, soothing, exciting, joyous, and graceful. A user study was conducted to test the synesthesia emotional image style transfer result with ground truth user perception triggered by the music-image pairs' stimuli. The transferred affective image result for music-image emotional synesthesia perception was proved effective according to user study result.

Keywords: Affective Computing, Image Style Transfer, Deep Convolutional Neural Networks

http://doi.org/10.3837/tiis.2021.04.015 **ISSN : 1976-7277**

1. Introduction

Synesthesia is known as an amazing cross-sensory phenomenon of human perception. We could sense the color shining from beautiful music melody, and hear the faint notes of music from artistic paintings. In an era with AI-driven multimedia innovation, there is little exchange between artists and scientists on the research of synesthesia intelligence. It is a challenging task to design computerized algorithm to create synesthesia style transfer on multimedia modalities. Here, driven by the request of cross-media matching and retrieval, music-image emotional synesthesia can be good research perspective of cross-media synesthesia style transfer exploration.

This work aims to generate emotional style images and considers the possible way of matching music with appropriate images in respect to emotional style. To date, not much work was conducted in this field. However, many successful image stylized transfer studies have achieved impressive results. Benefiting from the existing image transfer works, we proposed an emotion-driven image generator based on convolutional neural network through a loss combination exploration. The user study result indicates that the image generation result is satisfied on transferring the emotion character of reference music to images. The proposed network is light-weighted and efficient, and this strategy is flexible for application development.

We proposed the use of loss functions for training convolutional neural network for image stylization tasks. To sum up, the main contributions of this work can be concluded as follows:
(1) Our study demonstrates that music-image emotional cross-representation is an important part in multimedia transformation study, and satisfied emotional stylized image transformation result can be obtained by neural network;
(2) We proposed a novel emotion-driven image stylization framework for embedding the emotion style extracted from reference music, which creates an applicable way for music-image synesthesia emotional style transfer. The stylization network is established with VGG-19 network. In this way, we explored the way of realizing music-image cross-media transfer possible for application development.
(3) In order to manifest if the emotional representation of images can convey the emotional style of reference music, a user study experiment was conducted to verify the effectiveness of the proposed method.

The remainder of this paper is arranged as follows: Section 2 provides a review of the related works and methods in literature; Section 3 introduces the methods of music emotion recognition and image emotional style transfer; Section 4 presents the specific experiment procedure, including analysis of experimental results and user study for result evaluation; and Section 5 provides a conclusion and opportunities for further study. The general research roadmap of music-image synesthesia emotional style transfer study is presented in **Fig. 1**.

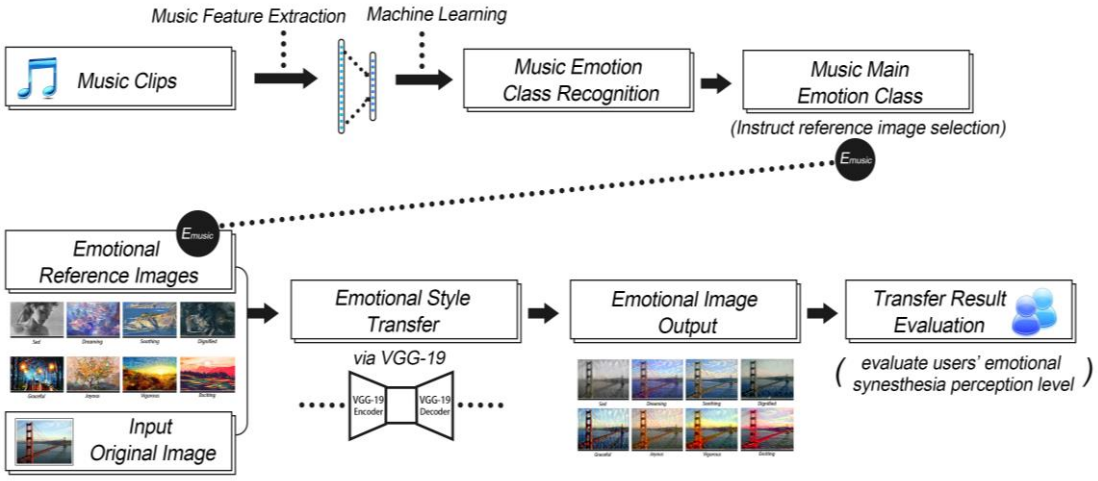

Fig. 1. Research flowchart of music-image synesthesia emotional style transfer study

2. Related Work

As stated in the related works, image style transfer has several crucial issues in the procedure: (1) image feature extraction; (2) image style representation; (3) style similarity measurement between the generated image and the reference image with structure loss function setting. According to our review, deep neural networks were widely used in image feature extraction in stylization, and CNN models were applied in style transfer stage. And the loss of local-structure and global structure will influence the detailed effect of stylized image output. A general review of existing studies with specific information of methods, experiment, functions, and results are presented in this section.

2.1 Image style transfer studies

Image style transfer is utilized for a wide variety of tasks, including image transformation, fashion style transfer, font style transfer and video style transfer. The main tasks in image style transfer can be concluded in two aspects: (1) image feature extraction with optimal method and network construction; (2) transfer performance generation and optimization exploration. In the existing literature, deep convolutional neural networks is constantly applied in image features extraction step. Image style synthesis performance were optimized based on loss function exploration and style representation algorithm design. Besides, the transfer performance is assessed by several indexes according to different senarios, including SSIM (Structural Similarity), IoU (Intersection-over-Union), inception-score, pixel-accuracy and etc.

Firstly, in the image feature extraction stage, deep convolutional neural network has achieved great performance and has potential in image style transfer. For instance, J. Johnson et al. developed a real-time style transfer method based on deep learning, using perceptual loss functions for training feed-forward networks for image transformation task. The method produces similar qualitative results with better efficiency. It achieves a loss comparable to 50~100 iterations of the baseline method [1]. M. Guo et al. proposed a spatial transfer strategy

for headshot portraits style transfer. They used VGG-19 to extract features and a deep spatially-aligned style transfer network was introduced. The result robustness outperforms existing benchmark methods [2]. B. Kim et al. used segment synthesis in fashion cloth style transfer based on Convolutional Neural Network (CNN). Three datasets were implemented in the experiment, including the Clothing Co-Parsing Dataset, Colorful-Fashion Parsing Dataset and Fashionista Dataset. The generator with pixel-wise training loss and feature reconstruction loss achieves the optimal performance, obtaining a highest style generation score, which shows superiority over benchmark methods [3]. Z. Lian et al. built a font style learning-based system to create large-scale handwriting fonts. Facing the difficulty of creating personal handwriting fonts, the researchers proposed "Easy Font" to learn the style from small dataset of personal handwriting characters. They employed Apply Neural Network to learn the overall handwriting style. Then a non-rigid point set registration method was used to set the correspondence relationship between target font and reference handwriting. The generation result was tested via Turing test of 97 participants, which indicates the effectiveness of the proposed method [4]. D.G. Aliaga et al. described an interactive building style transfer system based on simple user input. They applied a weighted combination of criteria to determine the fitness of a generated new face, including corner orientation, size and resolution [5]. L. Zhan et al. adopted zigzag learning method on CNN model parameters to realize fast neural style transfer effect on paintings [6]. H. Kwon et al. proposed a Deep Neural Network style-transfer learning method to generate CAPTCHA image with recognizable ability to users [7]. A. Khan et al. proposed a multi-convolutional-learning method to generate photo painting style with no restriction of specific style [8]. O. Jamriška et al. proposed a novel example-based video stylization approach. It can facilitate video sequence stylization by one or more key frames that users select to stylize with digital media [9]. R. Novak et al. improved an existing image artistic style transfer algorithm in their study. The improved method decreases the impact of visual artifacts and undesirable textures with activation shift, augmenting the style representation by using 16 layers and geometric weighting scheme to soften the style separation in repainting [10]. In 2015, L. A. Gatys generated artistic images on the basis of VGG-19. They found that replacing the max-pooling leads to improvement in gradient flow and more visual appealing result [11]. They also proposed an image artistic style synthesis method based on VGG neural network, which can separate and recombine the content and style of natural images [12]. Y. Shin et al. introduced a novel approach to undistorted faces in wide-angle shots photos, using a new energy minimization method. The proposed method can correct the unnatural distortions including stretching, squishing and skewing [13]. D. Guo et al. introduced an automatic face makeup method by using a makeup face as the style reference. Specifically, they decomposed the images into structural layers of face layer, skin detail layer and color layers for corresponding information transfer [14]. Seeing that the style textures tend to separate the output images, M. Cheng et al. explored a novel method with an additional structure representation. The global structure (represented by the depth map) and local structure details (represented by the image edges) are included in the new approach. The generation representation is impressive in objects depths presentation and dominant object clearly exhibition [15]. X. Zhang et al. developed a semantic correspondence guided deep image style transfer method to maintain the integrity of semantic structure in representation after color is migrated. The deep feature maps are extracted by VGG-19. A matting optimization technique is applied to ensure the semantic accuracy and representation faithfulness [16]. Consequently, convolutional neural network has great potential in tackling image style transfer tasks. H. Huang et al. used a feed-forward network for video stylization with temporally consistency. The algorithm employs a hybrid loss to extract the content of

input video frames, the style feature of the example style image and the temporal features of consecutive frames [17].

Secondly, image transfer result can be generated and optimized with different methods. H. Wu et al. found the orientation of each painting stroke can be the direction of generating a more natural and vivid stylized image. Their method has two stages, including direction-aware neural style transfer and texture enhancement [18]. Y. Gao et al. developed a new model AGIS-Net to implement shape and texture artistic style transfer on Chinese glyph image with small amount of input samples. The proposed method is consisted of several key factors, including content and style encoders, shape and texture collaborative working decoders and a local texture refinement loss for synthesis improvement [19]. C. Zhou et al. applied an improved feed-forward neural network for image style transfer with short running time. The image features were extracted by VGG-19. And gray conversion is conducted in data processing. The network contains two parts, including style transfer network and loss network. The method has good efficiency, and it can separate crucial target from the background to avoid the content loss [20]. For high-resolution images, problems of memory usage and time-consumption are still concerns for image stylization tasks. Z. Li et al. introduced a novel method to accelerate the speed of operation and reduce the memory usage with super-resolution style transfer network (SRSTN). Specifically, they first train the SRSTN with content loss only, and secondly they fine-tune the SRSTN with the pre-trained model with content loss and style loss [21]. D. Liang et al. conducted a study on robot calligraphy writing method exploration. In their work, input character images are transformed into stylized font via GAN method. Then the parameters of the stylized character images are extracted to instruct the robot writing operation. The final robot writing effect is good according to the experiment result [22]. F. Luan et al. proposed an approach, which can produce image stylization with local affine in color space, and it expresses the constraint as a custom differentiable energy parameter [23]. Y. Zhou et al. proposed BranchGAN for image-to –image transfer. Three factors were considered in the method, including pixel-level overall style, region semantics and domain distinguish ability, which is represented by reconstruction loss, encoding loss, and adversarial loss accordingly [24]. Fashion style transfer is an interesting application in image style transfer study. Tradition methods usually blend or warp the target clothes for the reference person. Y. Liu conducted a different approach namely SwapGAN to realize person-to-person fashion style transfer, which contains three generators and one discriminator [25]. J. Zhu et al. proposed an image-to-image translation in the absence of paired examples using cycle-consistent adversarial network. In the network structure, they applied GAN and PatchGAN for discriminator networks. The negative log likelihood objective is replaced by a least-squared loss in training procedure [26].

The specific information of the existing studies, including method, selected style reference images, applied datasets, research goal and experimental results are concluded in **Table 1**, in order to exhibit a clear overview of the related works.

Table 1. Image style transfer studies

Refer	Method	Style reference	Dataset	Research goal and results
J. Johnson, 2016 [1]	Propose the use of perceptual loss functions for training feed-forward networks	The Muse by Pablo Picasso	Microsoft-COCO dataset	It gives visually pleasing results with faster speed.
M. Guo, 2019 [2]	A robust spatial transformation strategy with pre-trained CNN	Spatially-transformed images	Headshot portraits	Achieve significant level of robustness comparing to benchmark result.

B. Kim, 2020 [3]	Segment synthesis with neural network	Clothing style images	Fashion-parsing datasets	Generate a new clothing styles images.
Z. Lian et al. [4]	Apply Neural Network in handwriting style learning; a non-rigid point set registration method for correspondence relationship	Handwriting style	Personal handwriting fonts dataset	To create large-scale handwriting fonts.
D.G. Aliaga et al. [5]	A weighted combination of criteria to determine the fitness of a architecture generation	Building style	-	Develop an interactive building style transfer system.
L. Zhan et al. [6]	Fast neural style transfer with Zigzag learning method on CNN parameters	Painting: Deux Femmes	-	Improve the stability of the method proposed by Gaty et al. [12] and achieves higher SSIM scores.
H. Kwon et al. [7]	DNN style-transfer learning method	Different CAPTCHA styles	Six CAPTCHA dataset in use on actual websites	CAPTCHA image generation that will resist recognition by machines while still has better recognizability to users.
A. Khan et al. [8]	Multi-convolutional-learning method for generation with no restriction of specific style	Stylized portrait paintings	Portrait paintings dataset	Propose a first approach for single-example photographic painting style in the wild.
O. Jamriška et al. [9]	A novel example-based video stylization approach	Stylized videos	-	Temporally coherent artistic stylization of video.
R. Novak et al. [10]	VGG-19, Geometric weighting scheme	Stylized images	-	Improved method decreases the impact of visual artifacts and undesirable textures.
In 2015, L. A. Gatys [11][12]	VGG-19, gradient descent, Gram matrix	Well-known artworks	-	They found that replacing the max-pooling by average pooling leads to improvement in gradient flow and better visual result.
Y. Shin et al. [13]	A new energy minimization method	-	-	Correct the unnatural distortions including stretching, squishing and skewing in wide-angle shots photos.
D. Guo et al. [14]	Decompose the images into structural layers.	Makeup face image	-	An automatic face makeup method.
M. Cheng et al. [15]	VGG-16	Stylized images	Microsoft COCO	The method achieves an impressive visual effectiveness. $SSIM_{content}$=0.629.
X. Zhang et al. [16]	VGG-19, nearest-neighbor field search	Stylized images	Portrait and scenery images	Maintain the integrity of semantic structure in representation.
H. Huang et al. [17]	Feed-forward network for style network, VGG-19 for loss network	Six exemplar styles : Gothic, Candy, Dream, Mosaic, Composition, Starry night	100 videos from Video.net	The result of proposed method receives higher user votes in user preference study.
H. Wu et al. [18]	VGG-16 for perceptual loss, direction filed loss is considered	Well-known artworks	-	Higher votes in user preference study.
Y. Gao et al. [19]	GANs Local texture refinement loss	Stylized fonts	English glygh image dataset Chinese artistic glyph	The result demonstrates the superiority of generating high-quality stylized glyph images. SSIM=0.6116; pixel-accuracy=0.7035.

			image dataset of 1571940 images	
C. Zhou et al. [20]	Feed-forword neural network VGG-19	Well-known artworks	-	The propose method can separate important targets from the background and realize style transfer in good efficiency.
Z. Li et al. [21]	Super-resolution style transfer network VGG-16 for loss network	Stylized images	Microsoft COCO	To solve the problems of memory usage and time-consumption in high-resolution image transfer. It produces competitive quality images with fast speed and less memory usage.
D. Liang et al. [22]	GAN based on deep learning	Stylized Chinese characters	-	Develop a robot calligraphy writing method. Similarity between the styles of robot writing characters and target characters is larger than 0.75.
F. Luan et al. [23]	VGG-19 as feature extractor, neural network as style network, DilatedNet for segmenting input image and reference image	Stylized images	-	User study result demonstrates the proposed approach produces photorealistic and faithful results.
Y. Zhou et al. [24]	BranchGAN	Photo labels image - image	Cityscapes	FCN-score: pixel-accuracy=78%, classification accuracy=29.7%, Intersection-over-Union=22.3%.
Y. Liu et al. [25]	SwapGAN	Persons images-clothing images	DeepFashion	For clothing swapping task. The result achieves a inception score=2.65+-0.09, SSIM=0.717.
J. Zhu et al. [26]	Cycle-consistent Adversarial Networks based on GAN	Map-aerial photo; labels-photo	Google maps; Cityscapes	Tackle the unpaired setting. The result is evaluated by AMT perceptual studies for realism perception (real) and FCN scores (highest in the comparison of benchmark methods).

(*SSIM: structural similarity index; DNN: Deep Neural Networks; CNN: Convolutional Neural Networks; FCN-score: Fully Convolutional Networks; AMT: Amazon Mechanical Turk.*)

2.2 Audio and video style transfer studies

Style transfer learning has also been widely studied in other multimedia fields. The related works can provide useful methods and inspiration for image style transfer. J. Yaniv et al. proposed a facial landmark detection algorithm based on neural network and presented some potential application that can apply the method, including geometry-aware style transfer, artist's geometric style generation and style signature [27]. Image style transfer technique is also applied in biomedical image research area. P. Andreini et al. presented a novel method for agar plate image segmentation based on deep convolutional neural network and GAN. The generator is scalable and it addresses the scarcity of biomedical data [28]. Z. Zhong et al. used a style transfer model to the challenging task of one view learning and unsupervised domain adaption in person re-identification [29]. Style transfer is also widely used in audio representation area, inspiring various applications and studies [30]. J. Chen et al. employed CNN for audio stylization to generate a stylized audio from an input audio [31]. K. Zsolnai-Feher et al. presented a material synthesis system, which can learn user preference via Gaussian Process Regression and create samples for recommendation [32]. M. Ruder et al. proposed a video artistic style transfer method based, which outperforms baselines both qualitatively and quantitatively. In the experiment, six famous paintings were utilized as reference style. They transfer the style from one artistic painting to a whole video. By applying a new initializations and loss functions applicable to videos, it generates stable stylized video sequences [33].

In conclusion, image artistic style transfer has been well discussed in existing studies. However, image representation in respect to emotional character and image transfer instructed by music emotion as a reference has not been well studied. Consequently, benefiting by the well study of music emotion recognition and image style transfer, we can propose a comprehensive and promising research scheme to realize music-image synesthesia emotional style transfer.

3. Proposed Approach

The music-image synesthesia emotional style transfer is aiming to produce image representation with emotional similarity to the reference music and integrity of original image content. Music-image synesthesia emotional style transfer can be framed as style transfer task with a loss function. As illustrated in **Fig. 2**, music-image synesthesia emotional style transfer procedure is composed of several sections: (1) music emotion recognition module; (2) reference emotional image selection; (3) image emotional style representation network; (4) image emotional stylization generator network.

In the representation network, emotional character can be expressed by image color, style and edge. Thus the combination of loss is used to realize emotional transfer, including style loss, color loss and edge loss. We applied convolutional neural network as the style transfer network in this experiment [10, 12]. The input image was transfer to be the output image in a style of specific emotion class recognized in the reference music piece. The image-to-image emotional style mapping is defined by a weighted combination of loss functions.

Specifically, firstly, the emotion class recognized in music is set as the emotional style target. Then, in the transfer training process, the input image is fed into the transform network to generate the stylized image output, while the input image remains to be the content reference. The transformation effect was measured by the general loss produced by the network, which is indicating the difference between generated image and content reference.

3.1 Image emotional transfer method

The main purpose of emotional style transfer is to apply the style of one reference emotional picture to another input image. The general experimental process is introduced as follows. Firstly, a white noise image is input into the network to generate a synthesis emotional style image with style transfer algorithm. Meanwhile, loss function is optimized by adjusting the pixel value of the input image, and the final output image is obtained with the minimum loss value as an optimal result.

The first step of style transfer algorithm is to feed style image and content image into VGG-19 network, and save the results of layers of conv1 ~ conv5. Different levels of the network can extract and learn the information of different levels of the image. The high-level features extracted by the network are describing information about object and layout of images, while the low-level features generally express the pixel information of images. Thus, different levels of abstract features are obtained, and the fusion of multi-layer features will enrich the style expression image style reconstruction in the following step. The satisfactory image synthesis result is that it retains a general content structure from the content image; besides, it also shows similar style character with the style reference image. The image style here refers to the image features of texture, color, visual character etc. The detailed transfer procedure is described in **Fig. 2**. And the specific network layers settings are illustrated in **Table 2**.

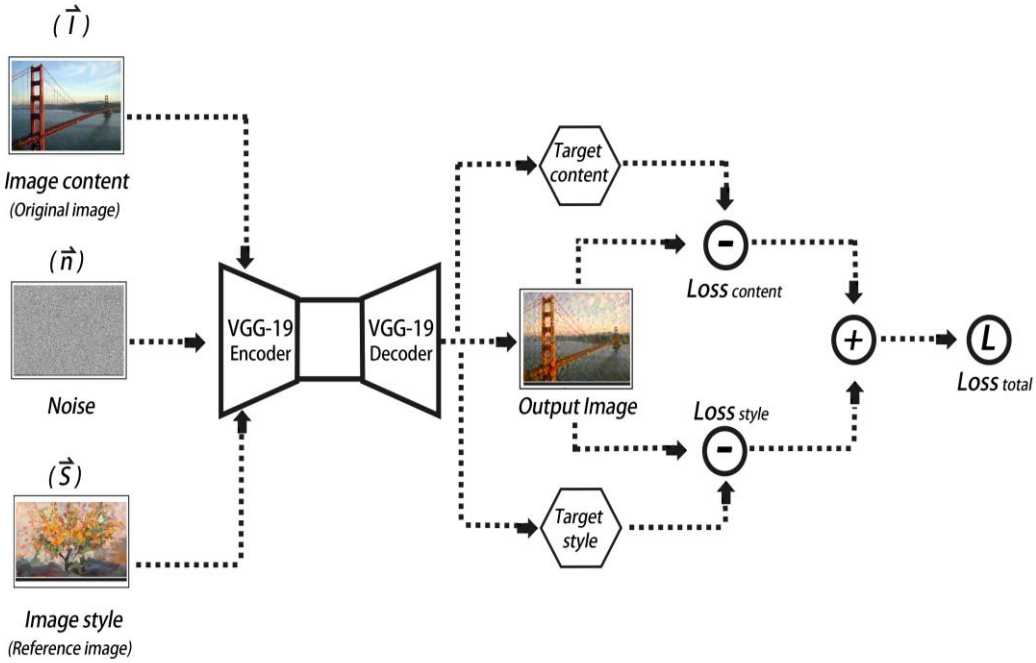

Fig. 2. Image transfer network

Table 2. Layers settings of image transfer network

Layers	Info
Content layers	Conv5_2
Style layers	Conv1_1, Conv2-1, Conv3-1, Conv4-1, Conv5-1
Weight schemes	Geometric: $w_k^s = 2^{Q-q(k)}$ $w_k^c = 2^{q(k)}$
Activation shift	S: The optimal result was achieved when the offset value s was set as -1. $\theta^k = \beta^k {\beta^k}^T$ $\theta^k = (\beta^k + s)(\beta^k + s)^T$ $\frac{\partial \theta^k}{\partial \beta^l} = 2(\beta^k + s)$
Blurred correlation	no
Adjacent activations correlation	yes: $\theta^{kb} = \beta^k [blur^{k-b} \circ up(\beta^b)]^T$

3.2 Loss computation

The total loss of the image synthesis is defined as follows:

$$L_{total}(\vec{\alpha},\vec{\beta},\vec{\varphi}) = aL_{content}(\vec{\alpha},\vec{\varphi}) + bL_{style}(\vec{\beta},\vec{\varphi}) \tag{1}$$

The above formula means that we hope that the style of the reference image is closer to that of the generated image, and the closer the content of the original image to the generated image, the better the result will be if the overall loss function is smaller. And the coefficient of a and b are introduced as weighting factors for content loss and style loss respectively.

(1) Loss of content

Image content mainly refers to its macro structure and outline. Rich global and abstract information can be extracted from the image with deep convolutional neural network. Therefore, the output of high-level activation function in convolutional neural network is used to define the content of the image. Then Euclidean distance is used as the index to measure the difference between content image and the generated image.

$$L_{content}(\vec{\alpha},\vec{\varphi},k) = \frac{1}{2}\sum_{i,j}(\beta_{ij}^{k} - P_{ij}^{k})^2 \tag{2}$$

Set the content image as P and the generated image as β in the k_{th} layer, so that the loss of content is calculated by least square method. Moreover, β_{ij} represents the j_{th} output value in the i_{th} feature map of the generated image.

$$\frac{\partial L_{content}}{\partial \beta_{ij}^{k}} = \begin{cases} (\beta^{k} - P^{k})_{ij} & if \quad \beta_{ij}^{k} > 0 \\ 0 & if \quad \beta_{ij}^{k} < 0 \end{cases} \tag{3}$$

Least square method is applied to get the minimum value, so that it can generate approximate images β close to the original picture P.

(2) Loss of style

In this experiment, we use Gram matrix in the same hidden layer to express the image style, since the Gram matrix can simulate the texture well and it has an excellent training effect.

$$G_{ij}^{k} = \sum_{w} D_{iw}^{k} D_{jw}^{k} \tag{4}$$

In which, G^k represents the Gram matrix corresponding to the response of the first layer, and D_{iw} represents the w_{th} elements of the response corresponding to the i_{th} convolution kernel of the layer. Therefore, each element of Gram matrix is to calculate an inner product, which describes a correlation between two responses, reflecting the style characteristics.

The next step is the same as the previous one. We take each layer of Gram matrix as a feature, making the Gram matrix of the reconstructed image as close as possible to the one of the original image for optimization.

$$\psi_{k} = \frac{1}{4N_{k}^{2}M_{k}^{2}}\sum_{i,j}(G_{ij}^{k} - \eta_{ij}^{k})^2 \tag{5}$$

Here least square method is used with a coefficient, where ψ represents a loss of this layer; *N*N* describes the size of matrix *G*, *M*M* is the size of *D*. And *G* represents the generated matrix, while η represents the matrix of the style image in this layer.

$$\frac{\partial \psi_k}{\partial D_{ij}^k} = \begin{cases} \frac{1}{N_l^2 M_l^2} ((D^k)^T (G^k - \eta^k))_{ij} & if \quad D_{ij}^k > 0 \\ 0 & if \quad D_{ij}^k < 0 \end{cases} \tag{6}$$

Similarly, the result is optimized though gradient descent method.

$$L_{style}(\vec{\beta}, \vec{\varphi}) = \sum_{k=0}^{K} \omega_k \psi_k \tag{7}$$

Each layer is weighted based on the previous method. The total loss of style will be obtained. The coefficient of *a* and *b* can be adjusted artificially.

To be concluded, we applied the neural algorithm of image style transfer method based on the original method proposed in [10, 11]. The experimental result turns out to be effective regarding to the appropriate emotion expressed in the output images. The major characters of the applied method are described as follows. Firstly, a per-layer content/style weighting scheme was applied; secondly, the method implements multiple layers to capture more style properties; thirdly, it uses shifted activations when computing Gram matrices to eliminate sparsity and make individual entries rich informative, and it also speed-up style transfer convergence; finally, targeting correlations of features of different layers was conducted to capture feature interactions.

4. Experiments

When music and image have the same emotional tendency, the emotional impact of the two senses will be more consistent. Nowadays, there are music based style transfer and image-based style transfer, which have good results. However, how to carry out cross sensory style transfer is a field rarely involved, and the related work is not rich. In this experiment, firstly, music emotion is identified. Based on the identified music emotion features, referring to the style characteristics of emotional pictures, the image is transformed into a certain emotional type style, so as to realize the work of music emotion guiding image style transfer. The experiment is carried out in two sections. The first section is designed to recognize music main emotional style category based on music feature analysis. In the second section of the experiment, music emotion class was adopted as reference to transfer the emotional style of the images and transferred emotional images were produced. The image emotional style transfer is implemented with Convolutional Neural Network. The experiment procedure can be concluded as follows, and also see **Fig. 3**:

***Step 1*. Emotional reference images preparation.** A total of 80 emotional images were collected from the internet by digital design and art major students. Twenty participants were invited to score the images based on Hevner emotion ring model. Then the image with the highest average rating score in each emotion class was selected to form the reference image data set for emotional style transfer experiment.

Step 2. **Emotional music preparation.** A Chinese folk emotional music database were utilized in this experiment. Music emotion recognition via machine learning method. Moreover, emotion class with the highest score in music emotion vector is identified as the main emotion of the music piece. The emotional style of the transfered image result is determined by this main emotion of the music, and the reference image in this emotion class will be selected accordingly.

Step 3. **Image emotional stylization based on convolutional neural network.**VGG-19 networks was applied in stylization to achieve the emotional stylized image. Please refer to section 3 for the specification of the networks.

Step 4. **Music-image synesthesia emotional style transfer performance evaluation via user study.** Twenty emotional music clips for each emotion classes. Thus, a total of 160 emotional music-image pairs were provided as synesthesia perception stimuli. Participants were asked to rate the 160 stimuli pairs on a scale of 0 to 5 to present the level of emotional synesthesia they sensed. Concequently, the image transfer performance was evaluated by their average rating score.

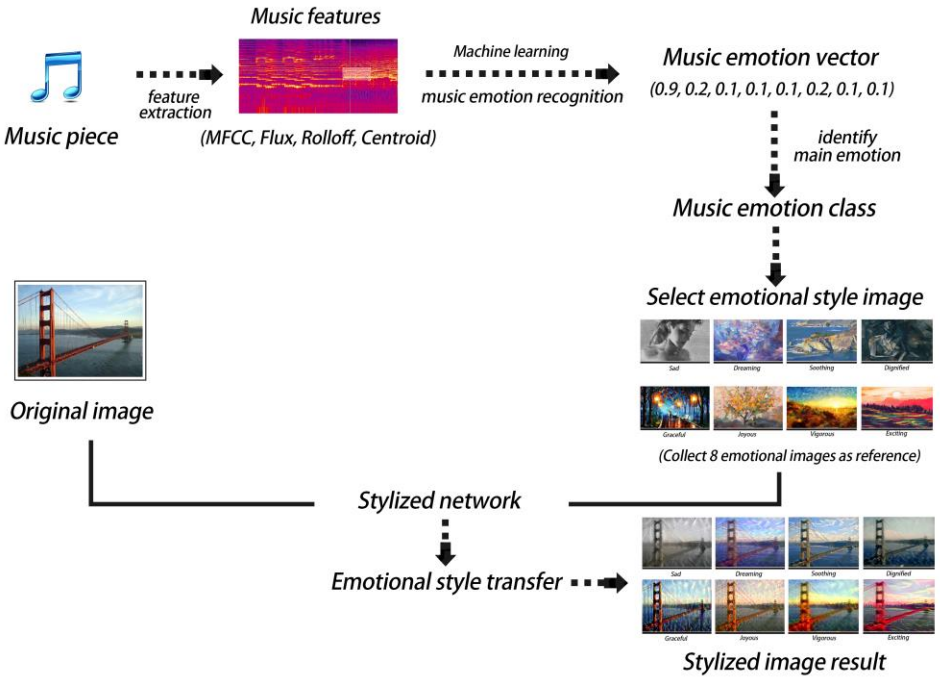

Fig. 3. Experimental procedure for music-image synesthesia emotional style transfer

4.1 Emotional reference images preparation

Emotional reference images for this experiment were collected from the internet. The reference image preparation process is presented as follows:

- Firstly, we invited five college students of digital design and art major to recommend 80 emotional images for the experiment. Each participant is asked to recommend two images for each emotion class. There are 10 images in each emotion category.
- Secondly, a total of 20 participants were invited to rate these emotional images in a scale of 0~1 for each emotion categories according to the emotion intensity they perceived.

Then each image can obtain an emotion vector $E_{emotion}$ in the rating experiment.

$$E_{emotion} = (E_{Dignified}, E_{Dreaming}, E_{Exciting}, E_{Graceful}, E_{Joyous}, E_{Sad}, E_{Soothing}, E_{Vigorous})$$

- Thirdly, the average score of the participants' ratings was set as the final emotion vector label.
- Finally, the image which obtains the highest score in each category was selected as the reference image with expressive emotion category for the emotional style transfer experiment, see **Fig. 4**. The emotion vector of the eight reference images were presented as follows:

Dignified image: $E_{emotion}$ =[**0.96**, 0.76, 0.02, 0.32, 0.01, 0.94, 0.23, 0.01]

Dreaming image: $E_{emotion}$ =[0.85, **0.97**, 0.66, 0.96, 0.92, 0.03, 0.88, 0.57]

Exciting image: $E_{emotion}$ =[0.67, 0.93, **0.95**, 0.89, 0.91, 0.02, 0.78, 0.92]

Graceful image: $E_{emotion}$ =[0.65, 0.89, 0.54, **0.91**, 0.75, 0.72, 0.65, 0.02]

Joyous image: $E_{emotion}$ =[0.26, 0.87, 0.45, 0.89, **0.91**, 0.01, 0.88, 0.63]

Sad image: $E_{emotion}$ =[0.78, 0.67, 0, 0.59, 0, **0.95**, 0.63, 0]

Soothing image: $E_{emotion}$ =[0.01, 0.92, 0.23, 0.93, 0.82, 0.02, **0.96**, 0.03]

Vigorous image: $E_{emotion}$ =[0.56, 0.9, 0.9, 0.84, 0.92, 0, 0.75, **0.93**]

Fig. 4. Emotional style reference images in eight emotion categories

4.2 Music emotion recognition

Music emotion can be recognized by machine learning method with great efficiency. In this experiment, we applied Hevner emotion ring model in music annotation, in order to realize music-image synesthesia emotional style transfer more specifically. There are eight emotion categories in Hevner emotion ring model, including Vigorous, Dignified, Sad, Dreaming, Soothing, Graceful, Joyous, and Exciting [34], see **Fig. 5**. A Chinese folk music emotion database with Hevner emotion labels were used for music-image matching, which contains 500 music pieces with a period of 30 seconds of a key emotional melody and a sampling rate of 16kHz. The specific information of database establishment can be found in the previous study of cross-media music-image emotional retrieval experiment [35].

Fig. 5. Hevner emotion ring model [35]

Features of MFCC; Centoid; Rolloff and Flux were extracted by OpenSmile to form the music feature database. Considering the possibility of developing music-image synesthesia emotional transfer system, recognition algorithms with good efficiency were selected for implementation, including libSVM (Support vector machine), RBFRegressor and RF (Random Forest) and Multilayer Perceptron. The music emotion recognition modeling was implemented with 10 folds cross validation.

Firstly, each class of eight emotions was recognized and the average recognition accuracy of eight classes was obtained as the final result. The detailed experimental results and model efficiency are presented in **Table 3** and **Table 4**. It is introduced that the optimal result was achieved by Random Forest. Random Forest method gets the result of an average RMSE (root mean squared error) of 0.1710 and an average CC (correlation coefficient) of 0.5004. Moreover, it also has great efficiency that it used 0.486 second in modeling. The performance of LibSVM is competitive. It used least average running time of 0.315 second in modeling, and it achieves an average RMSE of 0.1756 and an average CC of 0.5050.

Secondly, the recognized emotion class that has the highest value among all is defined as the main emotion expressed in the music. Consequently, the main emotion class will be utilized to select the corresponding reference image with the same emotion class for image style transfer in the next step of image style transfer.

Table 3. Music Emotion Recognition Test Result

Algor.	**RF**		**RBFRegressor**		**LibSVM**		**Multilayer Perceptron**	
	RMSE	CC	RMSE	CC	RMSE	CC	RMSE	CC
Dignified	0.1487	0.4676	0.1504	0.3954	0.1573	0.4723	0.1573	0.3954
Dreaming	0.1568	0.1703	0.1666	0.2581	0.156	0.1598	0.156	0.2581
Exciting	0.1803	0.7486	0.1827	0.7245	0.1797	0.7157	0.1797	0.7245
Graceful	0.1721	0.1434	0.1641	0.1912	0.1641	0.2615	0.1641	0.1912
Joyous	0.1797	0.703	0.173	0.643	0.1895	0.715	0.1895	0.643
Sad	0.1819	0.6398	0.1887	0.5536	0.1939	0.596	0.1939	0.5536
Soothing	0.1794	0.5067	0.181	0.4024	0.1907	0.4971	0.1907	0.4024
Vigorous	0.1693	0.6238	0.168	0.5874	0.1733	0.6223	0.1733	0.5874
Average	**0.1710**	**0.5004**	0.1718	0.4695	0.1756	**0.5050**	0.1756	0.4695

Table 4. Time Cost of Music Emotion Recognition Experiment (second)

Algor.	RF	RBFRegressor	LibSVM	Multilayer Perceptron
Dignified	0.68	0.87	0.51	8.59
Dreaming	0.39	0.49	0.22	8.96
Exciting	0.48	0.64	0.23	8.65
Graceful	0.37	0.36	0.20	8.59
Joyous	0.36	0.39	0.26	12.98
Sad	0.34	0.35	0.23	8.51
Soothing	0.33	0.18	0.26	8.77
Vigorous	0.94	0.64	0.61	9.01
Average	0.486	0.490	**0.315**	9.2575

4.3 Image emotional style transfer

Image emotional style transfer method using VGG-19 network was implemented in the experiment. We conducted experiments mainly on two tasks: image style transfer and network parameter optimization. Emotion category expressed in the music piece was used to select the corresponding emotional reference image, please see **Fig. 4**. The input image applied in this experiment is presented in **Fig. 6**.

The style transfer method was optimized to generate emotional image synthesis to match the music and set off the affective experience. And the total loss of content and style in transfer process was used as the evaluator in network performance.

The experiment procedure was introduced as follows:

Resize: The original images (content images) and emotional style reference images were resized as a scale of 250 x 250 pixels.

Fig. 6. Input image used in the experiment (Content image)

Training:

- The original images were optimized by a total of 300 iterations of L-BFGS algorithm. A batch size of 4 for 300 iterations was applied in the training, giving roughly 100 epochs over the training data. Adam was applied with a learning rate of 0.001.
- The optimization converges to optimal results can be achieved within 300 iterations. However, the efficiency of this method is relatively slow, due to the fact that each iteration requires a loss computation through VGG-19 network.

Synthesizing:

- The stylized images were synthesized and are regularized with total variation regularization.
- Specifically, the configuration of adjacent activations correlation was implemented over images in transfer, making an improvement based on blurred correlation in classical method, which can achieve a satisfactory output synthesis result.
- Parameter optimization:

Firstly, we held an experiment to seek for the optimal parameter a for the loss of content in (1). We tried to generate the least content with most style image when $a = 0.001$, and reconstruct the most content image with least style when $a = 10^3$. As shown in **Fig. 7**, a transition between content-similarity and style-similarity for Graceful emotional image can be seen by changing a from 0.001 to 10^3. A too small a value cannot prevent distortions, and thus the results have a non-photorealistic look. While, a too large a value suppresses the style in generating output image. Yielding a balanced look, we found the best parameter $a = 0.025$ to be the sweet spot to generate the result.

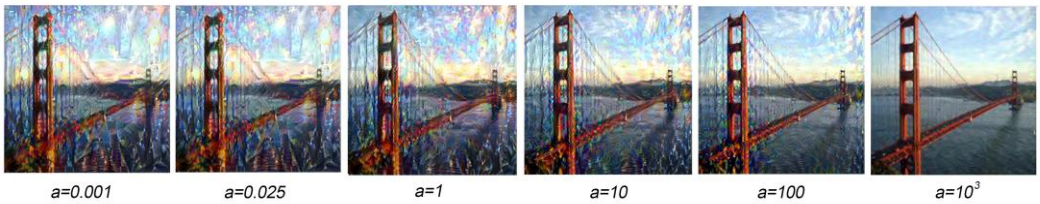

Fig. 7. Content-style trade-off for emotion style of Graceful. We control the balance between image content and style by giving different value to a in (1).

Secondly, in order to transfer emotion style to the output image, parameter b for loss of style in (1) was set according to the emotion vector $E_{emotion}$ of the reference image. For instance, in the image style transfer of Graceful emotion, b is defined as follows:

$b = E_{emotion}$ =[0.65, 0.89, 0.54, 0.91, 0.75, 0.72, 0.65, 0.02]

- We compared two kinds of loss computation method:
  $Loss_a$=1 / (2 * sqrt(channels) * sqrt(width * height));
  $loss_b$ =1 / (channels * width * height).
  Similar synthesis result is obtained, and the transferred image is shown in **Fig. 8**.

Fig. 8. Comparision of synthesis result of different loss computation method

- The loss of content reconstruction was got at layer relu2_2, and the loss of style reconstruction was computed at layers relu1_2, relu2_2, relu3_3, and relu4_3 of the VGG-19 network. The experiment was processes using a PC with Intel i7 3.40GHz and NVIDIA GeForce GTX 2080Ti.
- Lastly, in **Fig. 9** we presented synthesis result of the experiment.

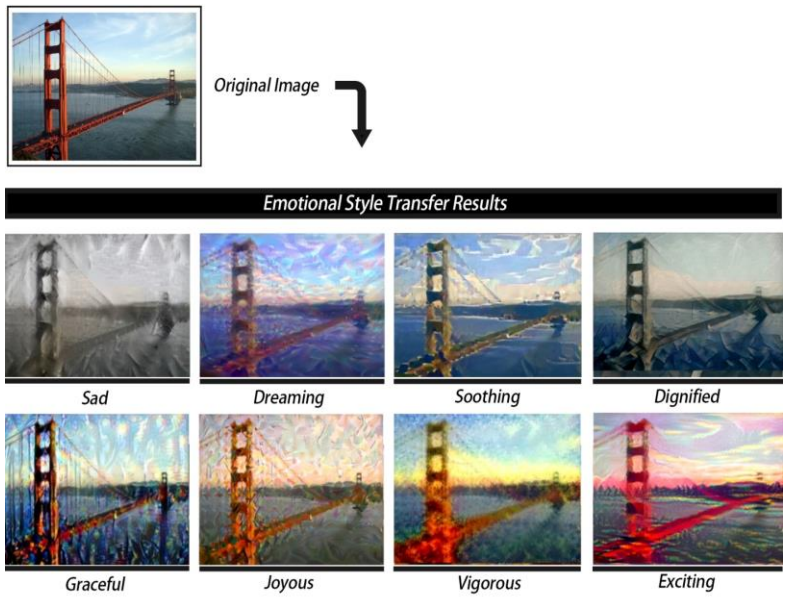

Fig. 9. Image emotional transfer results obtained by our solution

4.4 Music-image synesthesia emotional style transfer result evaluation

There are several evaluation indexes for evaluating image style transfer result in the existing research, including SSIM [6, 15, 19, 25] and pixel-accuracy [19, 24]. Inception-score [25] and IOU [24] were utilized in image style transfer with specific network or task. The specific information of these indexes is introduced as follows:

- SSIM(Structural Similarity): SSIM is an index for image quality assessment based on the comparison of three features, including luminance, contrast and structure. It is proposed to quantify the visibility of errors between a distorted image and a reference image. SSIM index has the advantages of simple and efficient.
- Pixel-accuracy: It is defined as the proportion of the number of correctly marked pixels to the total number of pixels.

It is thus clear that they are not quite suitable for emotional style transfer evaluation, since these indexes mainly assess the image quality and pixel accuracy between the generated image and the reference image. Therefore, user evaluation method was employed for assessing the effect of emotional style transfer result for music-image emotional synesthesia perception.

Consequently, a user study was conducted to test the synesthesia emotional image style transfer result with ground truth user perception evidence. Emotional music-image pairs were provided as synesthesia perception stimuli. Then we evaluate the level of participants' emotional synesthesia perception triggered by the music-image pairs' stimuli. A total of 20 participants (aged from 21 to 42 years; 10 female and 10 male) were invited to take part in the experiment. The experimental procedure is described as follows:

Step 1: **Music-image synesthesia emotional stimuli preparation**

We randomly selected the 20 emotional music clips for each emotional stylized image of the eight emotion classes presented in **Fig. 4** to evaluate the accordance in emotional similarity. A total of 160 emotional music-image pairs were provided as synesthesia perception stimuli.

Step 2: **Music-image synesthesia emotional stimuli presentation**

Each synesthesia emotional stimulus pair was presented on the PC screen (27 inches) for 10 seconds for perception. Then the participants will have a 5-second break before observing the next stimuli.

Step 3: **The rating of emotional synesthesia perception level**

Participants were asked to rate a group of 160 stimuli pairs on a scale of 0 to 5 to present the level of emotional synesthesia they sensed. A higher rating represents a higher level in synesthesia emotional perception, which also indicates satisfactory result in synesthesia emotional style transfer experiment. The rating procedure takes about 40 minutes for 160 music-image pairs by each participant. Consequently, the annotation of each stimuli pair is set as the average score of participants' ratings for further statistical analysis. The analysis of the synesthesia emotional perception annotation is given in **Table 5**.

By analysis, it is shown that the overall average score of the positive rate is 3.40. And the medians score is 3.6, variance score is 0.70, the skewness value is -1.30, the standard deviation value is achieved as 0.84, the maximum rating score is 5.0 and the minimum rating score is 0.1. The distribution maps of the user ratings in each emotion category are illustrated in **Fig. 10**.

Table 5. Analysis of the scores in the annotation experiment (in a scale of 1 to 5)

Statistics	Average	Medians	Variance	Skewness	Max	Min	STD
Results	3.40	3.6	0.70	-1.30	5.0	0.1	0.84

(Max: Maximum; Min: Minimum; STD: Standard Deviation)

Fig. 10. Distribution map of music-image synesthesia emotional pairs' rating results

As a result, the proposed approach was tested with user perception rating experiment and the scoring results have indicated the effectiveness of this method. It can be seen from the ratings distribution map that the music-image pairs of Dreaming, Exciting, Joyous, Sad, Soothing and Vigorous are perceived with a higher score than the pairs of Dignified and Graceful. It might due to the limitation of dataset, that the music data is not balance in different emotion categories. Additionally, the transferred image of Dignified and Graceful has a relatively poor performance, which cannot trigger the emotional synesthesia perception, according to the feedback of the participants. We found that the specific quantitative emotional synesthesia evaluation index is required for assessment, and it will be further explored in the future research.

5. Conclusion and Directions for Future Work

There is rich emotional synesthesia experience in human's perception, and there is implicit cross-media connection between image and music data. In this study, emotion features were selected as the bridge for cross-media style transfer basis. We proposed an image style transfer method, which is instructed by the emotion style recognized from music pieces. Consequently, an emotional synesthesia image-music pairs can be generated as an output. In the experiment, an optimal CNN network was applied to generate the emotional stylized images. This study provided an interesting and effective approach for emotional image style transfer and cross-media synesthesia multimedia pairs' generation. The image style transfer result was validated by a user study, which has proved the effectiveness of the method.

Image emotional style transfer aims to transform emotional characteristics in image while preserving the content, which is an abstract and high-level style transfer. Specifically, the reference image selection is an important aspect in emotional style transfer. In this experiment, the reference image was selected with discrete emotion label, in order to ensure the emotional style transfer can achieve satisfied results with significant emotion expression.

This study shows that it is possible to disentangle emotional features using an encoder-decoder network conditioned on discrete representation. The proposed networks make use of a VGG-19 network with loss computation to transfer emotional style during training. In this way, the network is able to transfer the emotional style of a reference image to a new image presentation. The current stylized results are not all satisfied. It might because that the choice of emotional style map is not accurate enough. A reference emotional image usually expresses different emotions and it can be understood in different way with personal experience. Consequently, the main drawback of the approach is that it is not able to create images with rich varieties of emotions. Since human perception for a piece of music or a painting is usually consisted of a series of emotions with different weights, a more advanced style transfer method should follow this rule to produce images with rich affective content, exhibit the complexity of affective synesthesia nature.

In the further study, music emotional style transfer method will be explored. A comprehensive binary emotional style transfer for image and music can construct a refined emotional synesthesia cross-media generation method for various scenarios. An application can be developed to generate synesthesia image-music pairs and provide image accompanied soundtrack creation or music accompanied stylized image re-creation.

Acknowledgements

This study is partly supported by the Natural Science Foundation of Zhejiang Province of China (LY19F020047) and National Natural Science Foundation of China (61402141).

References

[1] J. Johnson, A. Alahi, and F. Li, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *Proc. of European Conference on Computer Vision*, pp. 694-711, 2016. Article (CrossRef Link)

[2] M. Guo and J. Jiang, "A robust deep style transfer for headshot portraits," *Neurocomputing*, vol. 361, pp.164-172, Oct. 2019. Article (CrossRef Link)

[3] B. Kim, G. Kim, and S. Lee, "Style-controlled synthesis of clothing segments for fashion image manipulation," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 298-310, Feb. 2020. Article (CrossRef Link)

[4] Z. Lian, B. Zhao, X. Chen, and J. Xiao, "Easyfont: a style learning-based system to easily build your large-scale handwriting fonts," *ACM Transactions on Graphics*, vol. 38, no. 1, pp.1-18, Feb. 2019. Article (CrossRef Link)

[5] D. Aliaga, P. Rosen, and D. Bekins, "Style Grammars for Interactive Visualization of Architecture," *IEEE Transactions on visualization and computer graphics*, vol. 13, no. 4, pp. 786-798, July 2007. Article (CrossRef Link)

[6] L. Zhan and Y. Wang, "Stable and Refined Style Transfer Using Zigzag Learning Algorithm," *Neural Processing Letters*, vol. 50, pp. 2481-2492, Mar. 2019. Article (CrossRef Link)

[7] H. Kwon, H. Yoon, and K. Park, "CAPTCHA Image Generation: Two-Step Style-Transfer Learning in Deep Neural Networks," *Sensors*, vol. 20, no. 5, Mar. 2020. Article (CrossRef Link)

[8] A. Khan, M. Ahmad, N. Naqvi, F. Yousafzai, and J. Xiao, "Photographic painting style transfer using convolutional neural networks," *Multimedia Tools and Applications*, vol. 78, pp. 19565-19586, Feb. 2019. Article (CrossRef Link)

[9] O. Jamriška, Š. Sochorová, O. Texler, M. Lukáč, J. Fiser, J. Lu, E. Shechtman, and D. Sýkora, "Stylizing video by example," *ACM Transactions on Graphics*, vol. 38, no.4, pp. 1-11, July 2019. Article (CrossRef Link)

[10] R. Novak and Y. Nikulin, "Improving the neural algorithm of artistic style," *arXiv:1605.04603*, pp. 1-15, May 2016. Article (CrossRef Link)

[11] L. Gatys, A. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *Journal of Vision*, vol. 16, no. 12, pp. 1-16, Aug. 2016. Article (CrossRef Link)

[12] L. Gatys, A. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in *Proc. of IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, pp. 2414-2424, 2016. Article (CrossRef Link)

[13] Y. Shih, W. Lai, and C. Liang, "Distortion-free wide-angle portraits on camera phones," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1-12, July 2019. Article (CrossRef Link)

[14] D. Guo and T. Sim, "Digital face makeup by example," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 73-79, 2009. Article (CrossRef Link)

[15] M. Cheng, X. Liu, J. Wang, S. Lu, Y. Lai, and P. Rosin, "Structure-preserving neural style transfer," *IEEE Transactions on Image Processing*, vol. 29, pp. 909-920, Aug. 2019. Article (CrossRef Link)

[16] X. Zhang, X. Zhang, and Z. Xiao, "Deep photographic style transfer guided by semantic correspondence," *Multimedia Tools and Applications*, vol. 78, pp. 34649-34672, Dec. 2019. Article (CrossRef Link)

[17] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu, "Real-Time Neural Style Transfer for Videos," in *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, pp. 7044-7052, 2017. Article (CrossRef Link)

[18] H. Wu, Z. Sun, Y. Zhang, and Q. Li, "Direction-aware neural style transfer with texture enhancement," *Neurocomputing*, vol. 370, no. 22, pp. 39-55, Dec. 2019. Article (CrossRef Link)

[19] Y. Gao, Y. Guo, Z. Lian, M. Tang, and J. Xiao, "Artistic Glyph Image Synthesis via One-Stage Few-Shot Learning," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 1-12, Nov. 2019. Article (CrossRef Link)

[20] C. Zhou, Z. Gu, Y. Gao, and J. Wang, "An Improved Style Transfer Algorithm Using Feedforward Neural Network for Real Time Image Conversion," *Sustainability*, vol. 11, no. 20, Oct. 2019. Article (CrossRef Link)
[21] Z. Li, F. Zhou, L. Yang，X. Li , and J. Li, "Accelerate neural style transfer with super-resolution," *Multimedia Tools and Applications*, vol. 79, pp. 4347-4364, Feb. 2020. Article (CrossRef Link)
[22] D. Liang, D. Liang, S. Xing, P. Li, and X. Wu, "A robot calligraphy writing method based on style transferring algorithm and similarity evaluation," *Intelligent Service Robotics*, vol. 13, no. 1, pp. 137-146, Jan. 2020. Article (CrossRef Link)
[23] F. Luan, S. Paris, E. Shechtman, and K. Bala, "Deep Photo Style Transfer," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.6997-7005, 2017. Article (CrossRef Link)
[24] Y. Zhou, R. Jiang, X. Wu, J. He, S. Weng, and Q. Peng, "BranchGAN: Unsupervised Mutual Image-to-Image Transfer with A Single Encoder and Dual Decoders," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3136-3149, Dec. 2019. Article (CrossRef Link)
[25] Y. Liu, W. Chen, L. Liu, and M. Lew, "SwapGAN: A Multistage Generative Approach for Person-to-Person Fashion Style Transfer," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2209-2222, Sep. 2019. Article (CrossRef Link)
[26] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2242-2251, 2017. Article (CrossRef Link)
[27] J. Yaniv, Y. Newman, and A. Shamir, "The face of art: landmark detection and geometric style in portraits," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1-15, July 2019. Article (CrossRef Link)
[28] P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, and F. Scarselli, "Image generation by GAN and style transfer for agar plate image segmentation," *Computer Methods and Programs in Biomedicine*, vol. 184, no. 105268, Feb. 2020. Article (CrossRef Link)
[29] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "CamStyle: A Novel Data Augmentation Method for Person Re-Identification," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1176-1191, Mar. 2019. Article (CrossRef Link)
[30] M. Huzaifah and L. Wyse, "Applying Visual Domain Style Transfer and Texture Synthesis Techniques to Audio - Insights and Challenges," *Neural Computing and Applications*, vol. 32, no. 4 , pp. 1051-1065, Feb. 2019. Article (CrossRef Link)
[31] J. Chen, G. Yang, H. Zhao, and M. Ramasamy, "Audio style transfer using shallow convolutional networks and random filters," *Multimedia Tools and Applications*, vol. 79, no. 21-22, pp. 15043-15057, June 2020. Article (CrossRef Link)
[32] K. Zsolnai-Feher, P. Wonka, and M. Wimmer, "Gaussian Material Synthesis," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1-14, Aug. 2018. Article (CrossRef Link)
[33] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proc. of German Conference on Pattern Recognition*, pp. 26-36, 2016. Article (CrossRef Link)
[34] K. Hevner, "Experimental studies of the elements of expression in music," *The American Journal of Psychology*, vol. 48, no. 2, pp. 246-268, Apr. 1936. Article (CrossRef Link)
[35] B. Xing, K. Zhang, S. Sun, L. Zhang, Z. Gao, J. Wang, and S. Chen, "Emotion-driven Chinese folk music-image retrieval based on DE-SVM," *Neurocomoputing*, vol. 148, pp. 619-627, Jan. 2015. Article (CrossRef Link)

Baixi Xing received the Bachelor degree from Nanjing University of Aeronautics and Astronautics. She received the Ph.D. degree in Digital Art and Design from Zhejiang University, Hangzhou, China, in 2014. She worked as a Post-doctor in the College of Computer Science and Technology, Zhejiang University from 2015 to 2018. She is now an assistant professor in Institute of Industrial Design, Zhejiang University of Technology. Her research interest lies in affective computing and aesthetics computing. At present, she is focusing in multimodal emotion recognition and cross media retrieval. She is also interested in the research of human computer interaction and user experience design.

Jian Dou is a graduate student in School of Media and Design, Hangzhou Dianzi University. His research interest lies in the field of data mining, multimedia emotion analysis and music information retrieval.

Qing Huang is a graduate student in School of Computer Science and Technology, Hangzhou Dianzi University. His research interest lies in the field of image style transfer, multimedia information analysis and aesthetics computing.

Huahao Si is a graduate student in School of Media and Design, Hangzhou Dianzi University. His research interest lies in the field of affective computing, multimedia information analysis and music information retrieval.