ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095200

MSNET: A DEEP ARCHITECTURE USING MULTI-SENTIMENT SEMANTICS FOR SENTIMENT-AWARE IMAGE STYLE TRANSFER

Shikun Sun[1], *Jia Jia*[1,2], *Haozhe Wu*[1], *Zijie Ye*[1], *Junliang Xing*[1,*]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology

ABSTRACT

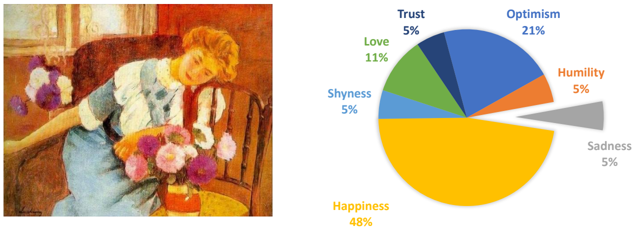
Sentiment plays an essential role in people's perception of images. To incorporate the sentiment information into the image style transfer task for better sentiment-aware performance, we introduce a new task named sentiment-aware image style transfer. To solve this problem, we first introduce a novel Multi-Sentiment Semantics Space (MSS-Space) to capture the non-deterministic and complicated nature of sentiment semantics. With the MSS-Space, we establish tight associations between the visual attributes of images and the multi-sentiment semantics by minimizing their distance in MSS-Space and then propose the Multi-Sentiment Style Transfer Net (MSNet). Experiments demonstrate that, compared with three competing models, our proposed MSNet generates more explicit images and better preserves the integrity of salient objects, local details, and multi-sentiment. In particular, our model outperforms the state-of-the-art by +28.72% in terms of the top-3 accuracy on average.

Index Terms— Style Transfer, Sentiment Analysis, self-attentive, Perceptual Loss

1. INTRODUCTION

Sentiment-aware image style transfer, which aims to take the sentiment of the artwork into account when transferring an image style, is a very challenging task in computer vision. As shown in Fig. 1, the images may reflect the complicated sentiment semantics of the creators, and the audiences of the images will also have different feelings about them [1]. These issues make the result of sentiment-aware image style transfer nondeterministic and dependent on multiple complicated factors.

Most previous work on image style transfer [2, 3, 4, 5, 6, 7] renders the style of an artwork to a realistic photo while preserving the content of the photo. Huang et al. [2] leverage AdaIN module to investigate style transfer first. Hu et al. [3] introduce aesthetics to this task. Karras et al.[6] use GAN to conduct style transfer first. However, since the sentiment semantics are complex and uncertain to describe [8], these works seldom consider the sentiment semantics. Even though Chen et al. [9] and An et al. [10] have researched image sentiment transfer, the domain of their referred sentiment image is realistic photos, and they only focus on the realistic output, which means that their task is different from image style transfer. Moreover, they ignore updating the description of the sentiment to support the intricate associations between visual attributes and multi-sentiment semantics. Therefore, sentiment-aware image style transfer remains a challenging and unexplored problem. This work aims to incorporate the sentiment semantics into the image style transfer task for more sentimentally meaningful results. To this end, we introduce a novel probability-based vector space, named Multi-Sentiment Semantics Space (MSS-Space), to depict the sentiment inspired by artworks in a more fine-grained manner. Each vector in the MSS-Space denotes a sentiment distribution inspired by an artwork, and each component of the vector signifies the probability that people perceive particular sentiment. Moreover, compared with two traditional manners of sentiment representation: Categorical Emotion States (CES) and Dimensional Emotion Space (DES) [11], our MSS-Space takes advantage of both sides. More specifically, CES data is easy to get but coarsely describes the sentiments, while DES data provides detailed sentiment representation but is hard to acquire. In contrast, MSS-Space data is easy to obtain and describes fine-grained sentiments.



Interior by Stefan luchian

The sentiment distribution

Fig. 1: An artwork and the corresponding sentiment distribution. Although the majority of people feel positive emotions from this artwork, there are still a small number of people who feel sadness.

By minimizing their distance in MSS-Space, we establish tight associations between the visual attributes of images and the multi-sentiment semantics. Then we proposed a new deep architecture Multi-Sentiment Transfer Net (MSNet). Additionally, to enforce the model to focus on sentiment-related objects, we designed a symmetric self-attentive neural network based on the work of [12] in the MSNet. Our main contributions are three-fold:

1. We introduce a novel sentiment-aware image style transfer task, which aims to consider the sentiment semantics of artworks for better sentiment-aware consistency between the output and the artwork.
2. We leverage the MSS-Space to represent the sentiment semantics, which has the advantages of both DES and CES: being easy to get and adequate expressiveness, more accuracy, and more diversity.
3. Based on MSS-Space, we propose a new deep architecture, referred to as MSNet, which comprehensively explores the association between visual attributes and sentiment semantics of images.

With the above technical contributions, we have obtained a high-performance sentiment-aware image style transfer model. Extensive experimental results demonstrate the effectiveness of the proposed

[*] Corresponding author

model. Specifically, our MSNet model outperforms state-of-the-art methods by +28.72% in terms of top-3 accuracy on average.

2. METHODOLOGY

To transfer the aesthetic style of the reference image to the content image while considering the image sentiment semantics, firstly we leverage the Multi-Sentiment Semantics Space (MSS-Space) to get a sentiment-aware perceptual loss $\mathcal{L}_{se}$. Then with the $\mathcal{L}_{se}$, we propose the Multi-Sentiment Semantics Style Transfer Net (MSNet), which has a novel self-attentive module. The overall framework of our proposed approach is illustrated in Fig 2.

2.1. Multi-Sentiment Semantic Space

Firstly, we construct the MSS-Space. We use the probability distribution of the sentiments the audiences perceive from a specific artwork to describe the sentiment semantics of the artwork. The probability vector of the probability distribution is the label of the artwork in the MSS-Space. Then, we train a neural net $\widehat{\mathrm{Ex}}_{se}$ to extract the sentiment probability vector from artworks. However, we do not directly use the Euclidean distance of the outputs of $\widehat{\mathrm{Ex}}_{se}$ as the sentiment-aware perceptual loss $\mathcal{L}_{se}$. The reasons are two-fold: firstly, the dimension of the MSS-Space depends on the number of categories of sentiments, which differs from dataset to dataset, and secondly, there are entangled relationships between different categories of sentiments that are difficult to capture. Instead, we use meta-MSS-Space to describe the disentangled sentiment semantics. The meta-MSS-Space denotes the space spanned by outputs from the penultimate layer of $\widehat{\mathrm{Ex}}_{se}$, whose dimension is fixed. The dimension of meta-MSS-Space is much larger than MSS-Space for better generalization. We note the network from image to meta-MSS-Space as Ex_{se}. The relationship between Ex_{se} and $\widehat{\mathrm{Ex}}_{se}$ is

$$\widehat{\mathrm{Ex}}_{se} = MLP\left(\mathrm{Ex}_{se}\right). \tag{1}$$

There are two advantages of the meta-MSS-Space:

- The meta-MSS-Space provides a more generalized description of the sentiment semantics, which is adaptable to different sentiment datasets.
- The meta-MSS-Space is a better-disentangled sentiment semantics space.

Then we use the Euclidean distance in the meta-MSS-Space as the sentiment-aware perceptual loss $\mathcal{L}_{se}$. The evaluation of the Multi-Sentiment Semantic depends on the following MSS-Loss:

$$\mathcal{L}_{se} = ||\mathrm{Ex}_{se}(p_{ij}) - \mathrm{Ex}_{se}(a_j)||_2. \tag{2}$$

2.2. Multi-Sentiment Transfer Net

2.2.1. Feature Extractor and Decoder

We use the encoder-decoder model as our backbone and mix the information of content and reference images during the encoding and decoding progress. We use a VGG-19 [13] with Batch Normalization [14] as the feature extractor and a symmetrical neural network as the decoder [15].

2.2.2. Residual Symmetrical Self-attentive Module

Attention module [16, 17] makes excellent progress in extensive researches. Based on the work of Yao et al. [12], we create a symmetrical attention neural network with fewer parameters to make the net more expressive on the content. In Reformer[18] they find this does not harm the performance. According to that, we use the same convolution kernel k_1 for Q and K :

$$Q = K = f_{i1}^{HW\times\frac{C}{2}} = p_{ih}^{HW\times C} \otimes k_1. \tag{3}$$

Then we use another convolution kernel k_2 for V:

$$V = f_{i2}^{HW\times C} = p_{ih}^{HW\times C} \otimes k_2. \tag{4}$$

After that we get the attention map:

$$att_i^{HW\times C} = \frac{2}{\pi}\arctan\left(\mathrm{softmax}(QK^T)V\right). \tag{5}$$

Finally, unsqueeze $att_i^{HW\times C}$ as $att_i^{H\times W\times C}$ to get the residuals of the factors of attention. Then we get the enhanced feature map:

$$\hat{p}_{ih}^{H\times W\times C} = p_{ih}^{H\times W\times C} + att_i^{H\times W\times C} \odot p_{ih}^{H\times W\times C}. \tag{6}$$

2.2.3. Mix of Feature Information

Now we get two feature maps, one of which is $\hat{p}_{ih}$ from original content picture p_i and the other is a_{jh} from a_j. Then we use the assumption that in this deep feature space, the means and variances of all pixels represent the sentiment of the image, and the other information describes the content of the picture. We use an AdaIN module to do the basic transfer:

$$\mathrm{AdaIN}(\hat{p}_{ih}, a_j) = \frac{\sigma(a_j)}{\sigma(\hat{p}_{ih})}\left(\hat{p}_{ih} - \mu(\hat{p}_{ih})\right) + \mu(a_j), \tag{7}$$

where σ and μ are the standard deviation and mean functions.

2.2.4. Basic Content and Style Losses

We follow the work of [2, 3, 12] and use the same content and style losses. Their job is to do image style transfer, and that job has effective two losses to train the decoder after the AdaIN module. One is content loss, and another is basic style loss. This technique uses the pre-trained VGG-19-BN (the same as the encoder) as the ”metric”. In more detail, they use the Euclidean distance between deep feature maps to measure the content similarity:

$$\mathcal{L}_{co} = \left\|\mathrm{Ex}_{cs}\left(p_{ij}\right) - \mathrm{Ex}_{cs}\left(p_i\right)\right\|_2. \tag{8}$$

Also, they use the Euclidean distance between $\mu s, \sigma s$ of feature maps to measure the style similarity:

$$\begin{aligned}\mathcal{L}_{st} = &\sum_{l=1}^{4}\left\|\mu\left(\mathrm{Ex}_{csl}(p_{ij})\right) - \mu(\mathrm{Ex}_{csl}(a_j))\right\|_2 \\ &+\sum_{l=1}^{4}\left\|\sigma(\mathrm{Ex}_{csl}(p_{ij})) - \sigma(\mathrm{Ex}_{csl}(a_j))\right\|_2.\end{aligned} \tag{9}$$

Note that the metric contains the distance of VGG’s four layers.

2.2.5. Metric of The Aesthetics

Besides, we also use some critical features in traditional machine learning methods to ensure the basic aesthetic features are not poles apart. We choose nine color features and twenty-four texture features. The color features are different order moments of different colors that contain nine components: 1st, 2nd, and 3rd order moments of red, green, and blue. And the texture features are mainly

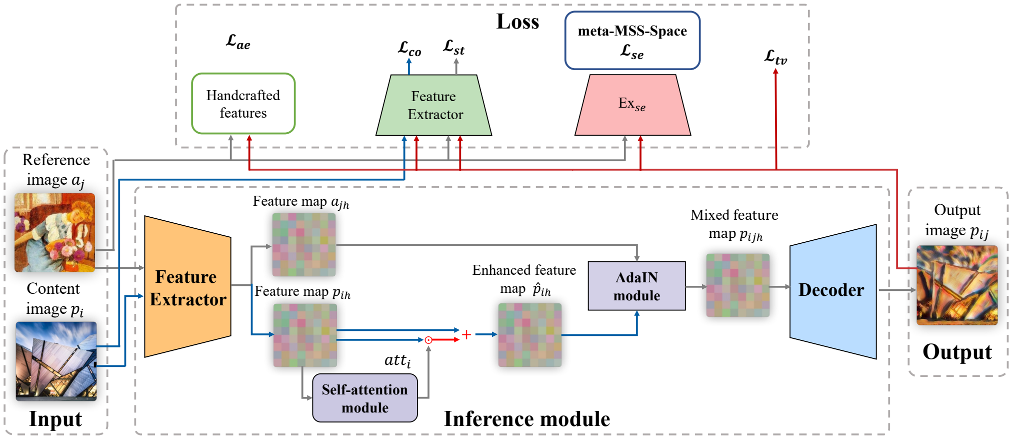


Fig. 2: Our model pipeline. The content image p_i and reference image a_j are fed into the same feature extractor to get their corresponding encodings as p_{ih} and a_{jh} respectively. After that, a self-attentive module is designed to get an attention map att_{ih} of p_{ih} and the enhanced feature map $\hat{p}_{ih}$. Then, AdaIN [2] module transfers the multi-sentiment transfer between feature maps. Finally, there is a decoder with inverse symmetry to the feature extractor to get the output image p_{ij}.

from Gray Level-gradient Co-occurrent Matrix [19] (GLCM), which contains 24 components: the contrast, the dissimilarity, the angular second moment (ASM), the energy, and the correlation in four directions (up, down, left, right). We have a similar loss of aesthetic features as the sentiment features:

$$\mathcal{L}_{ae} = \left\|\mathrm{Ex}_{ae}(p_{ij}) - \mathrm{Ex}_{ae}(a_j)\right\|_2 . \tag{10}$$

Note that some of the aesthetic features are not derivable. We use another neural network to fit the extraction of these features to solve this problem.

Besides, to prevent distortion of the generated images after the self-attentive mechanism and to reduce the high-frequency noise of the generated images, we use the total variance [20, 21] technique. For a two-dimensional continuously differentiable signal $f(u, v)$, the discrete total variance $\mathcal{R}_{V^\beta}(f)$ of order $\beta = 2$ is

$$\begin{aligned}\mathcal{L}_{tv} = &\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W-1}\left(p_{ij}^{h,w,c} - p_{ij}^{h,w+1,c}\right)^2 \\ +&\sum_{c=1}^{C}\sum_{h=1}^{H-1}\sum_{w=1}^{W}\left(p_{ij}^{h,w,c} - p_{ij}^{h+1,w,c}\right)^2 .\end{aligned} \tag{11}$$

To sum up, our loss is the affine combination of the losses introduced before:

$$\begin{aligned}\mathcal{L}_{total} =& \lambda_1\mathcal{L}_{co} + \lambda_2\mathcal{L}_{st} + \lambda_3\mathcal{L}_{se} \\ &+ \lambda_4\mathcal{L}_{ae} + \lambda_5\mathcal{L}_{tv}.\end{aligned} \tag{12}$$

3. EXPERIMENT

3.1. Dataset

For the content dataset, we choose the training set of COCO [22] which contains 82,783 real-world photos. For the artwork dataset, we use WikiArt Emotions [1] which contains 4,105 pieces of artwork and multi-sentiment labels. For example, the sentiment label of the artwork in Fig. 1 is the probabilities of sentiments people feel, which is a 20-dimensional vector, 7 of which are non-zero and the dimension representing happiness is the largest.

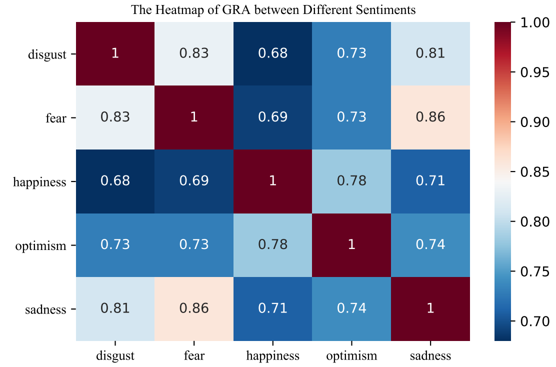


Fig. 3: The heatmap of GRA between five different sentiments. Zoom in for details.

As shown in Fig. 3, we dig the relationships between different sentiments by Grey Relational Analysis(GRA) [23]. We find that some sentiments rarely appear simultaneously, e.g., disgust and happiness, which means that some sentiments have conflicts.

3.2. Implementation Details

We train the $\widehat{\mathrm{Ex}}_{se}$ module first, and the dimension of meta-MSS-Space is 1000. We use 2,105 images in dataset A as training data while using another 2,000 images for validation. $\widehat{\mathrm{Ex}}_{se}$ module contains a pre-trained VGG net and a 3-layer MLP. Finally, the top-3 accuracy is about 73%. Then we connect the pre-trained VGG encoder and the decoder directly to train the decoder.

Methods	Top-1 Accuracy ↑	Top-3 Accuracy ↑
AdaIN [2]	6.09%	19.27 %
WCT [24]	7.12%	18.34%
StyleFlow [25]	8.35%	20.64%
AesUST [26]	8.64%	21.23%
MSNet (Ours)	**26.58%**	**49.95%**

Table 1: Quantitative comparison.

Methods	CDP ↑	TDP ↑	IP↑	DP ↑	MSP ↑
AdaIN [2]	21.48%	**47.14%**	45.71%	40.00%	21.43%
WCT [24]	**47.14%**	7.14%	4.29%	8.57%	27.14%
MSNet (Ours)	31.43%	45.72%	**50.00%**	**51.43%**	**51.43%**

Table 2: User study.

We use MSE as the self-supervision loss. The learning rate is 0.0001. The reconstruction result is shown in Fig. 4. After that, we start to train the whole pipeline with $\mathcal{L}_{total}$.

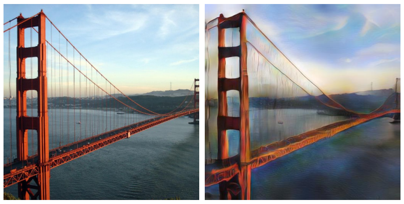


Fig. 4: Results of reconstruction.

3.3. Objective Measures

We use WikiArt Emotions training a sentiment classification function J_{se} to test the classification accuracy. We use 100 content pictures and 100 sentiment images to get 10,000 for the accuracy test with SOTA baselines. The result is shown in Table 1.

3.4. Ablation Study

To show the effectiveness of our self-attentive module, we do an ablation experiment. The result is as shown in Fig. 5. We relieve the problem of transferring the solid color blocks to strange textures with the self-attentive module, which indicates the effects of our attention module.

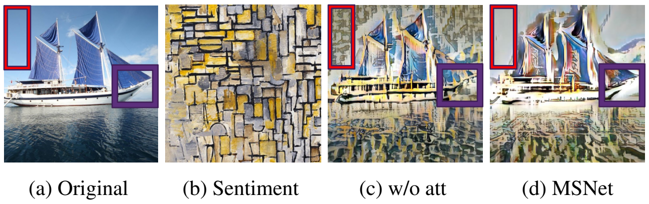


Fig. 5: The comparison between MSNet w/o att and MSNet.

3.5. Subjective Experiment

We compare the transfer effects of our MSNet with some other style transfer models, as shown in Fig. 6. We find that compared with other models, MSNet has the following advantages:

1. **MSNet preserves the main objects better.** See the last two rows of the Fig. 6. Thanks to the total variation loss $\mathcal{L}_{tv}$, We achieve a good balance of detail preservation and main object preservation.
2. **The clean areas are transferred cleanly.** See the fourth row of Fig. 6. With the power of our self-attentive module, The clean, blue sky is transferred cleanly with our model.
3. **The style transfer effects of our MSNet are comparable to other style transfer models.**

3.6. User Study

We conducted a user study to compare the perceptual quality of transfer results generated by various models. In this study, 15 users compared the results generated by different models from the same photo and artwork and selected the best one based on the following evaluation indicators:

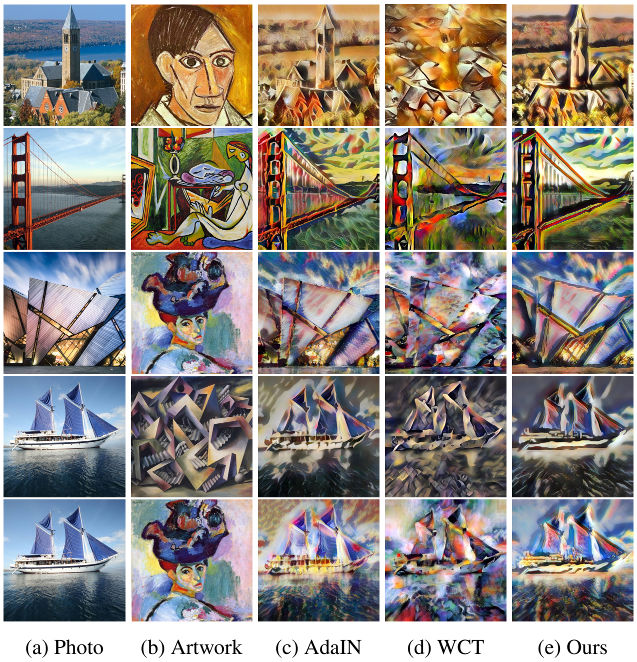


Fig. 6: The comparison of different models. Zoom in for details.

- **Color Distribution Preservation (CDP)**: The ability to preserve the color distribution of artwork.
- **Texture Distribution Preservation (TDP)**: The ability to preserve the texture distribution of photo. In other words, do not transfer strange textures on the clean color areas.
- **Integrity Preservation (IP)**: The ability to preserve the integrity of major objects in photo.
- **Detail Preservation (DP)**: The ability to preserve the details in photo.
- **Multi-Sentiment Preservation (MSP)**: The ability to preserve perceptual multi-sentiments triggered by artworks.

The percentage of preference is shown in Table 2. We find that MSNet performs best in IP, DP, and MSP, which indicates that our MSNet captures the main object and transfer the sentiment of artworks better.

4. CONCLUSION

This paper introduces a novel sentiment-aware image style transfer task. To solve this problem, firstly we build a Multi-Sentiment Semantics Space to describe sentiment more finely, and based on MSS-Space and residual symmetrical self-attentive module, we propose a novel MSNet to do style transfer with multi-sentiment semantics. Extensive experiments including both sentiment classifier evaluation and user rewards indicate that our model outperforms other state-of-the-art methods on sentiment-aware performance.

5. ACKNOWLEDGEMENTS

This work is supported by the National Key R&D Program of China under Grant No.2021QY1500, and the state key program of the National Natural Science Foundation of China (NSFC) (No.61831022).

6. REFERENCES

[1] Saif Mohammad and Svetlana Kiritchenko, “Wikiart emotions: An annotated dataset of emotions evoked by art,” in *International Conference on Language Resources and Evaluation*, 2018. 1, 3

[2] Xun Huang and Serge Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *International Conference on Computer Vision*, 2017. 1, 2, 3, 4

[3] Zhiyuan Hu, Jia Jia, Bei Liu, Yaohua Bu, and Jianlong Fu, “Aesthetic-aware image style transfer,” in *International Conference on Multimedia*, 2020. 1, 2

[4] Leon Gatys, Alexander Ecker, and Matthias Bethge, “A neural algorithm of artistic style,” *Journal of Vision*, vol. 16, no. 12, pp. 326–326, 2016. 1

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European Conference on Computer Vision*, 2016. 1

[6] Tero Karras, Samuli Laine, and Timo Aila, “A style-based generator architecture for generative adversarial networks,” in *Conference on Computer Vision and Pattern Recognition*, 2019. 1

[7] Anh-Duc Nguyen, Seonghwa Choi, Woojae Kim, and Sanghoon Lee, “A simple way of multimodal and arbitrary style transfer,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1752–1756. 1

[8] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas, “Artemis: Affective language for art,” *Computer Research Repository*, vol. abs/2101.07396, 2021. 1

[9] Tianlang Chen, Wei Xiong, Haitian Zheng, and Jiebo Luo, “Image sentiment transfer,” in *International Conference on Multimedia*, 2020. 1

[10] Jie An, Tianlang Chen, Songyang Zhang, and Jiebo Luo, “Global image sentiment transfer,” in *International Conference on Pattern Recognition*, 2021. 1

[11] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer, “Affective image content analysis: A comprehensive survey,” in *International Joint Conference on Artificial Intelligence*, 2018. 1

[12] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang, “Attention-aware multi-stroke style transfer,” in *Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[13] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014. 2

[14] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, 2015. 2

[15] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu, “Invertible image rescaling,” in *European Conference on Computer Vision*, 2020. 2

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Neural Information Processing Systems*, 2017. 2

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018. 2

[18] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya, “Reformer: The efficient transformer,” *arXiv preprint arXiv:2001.04451*, 2020. 2

[19] R.M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973. 3

[20] Leonid I Rudin, Stanley Osher, and Emad Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992. 3

[21] Aravindh Mahendran and Andrea Vedaldi, “Understanding deep image representations by inverting them,” in *Conference on Computer Vision and Pattern Recognition*, 2015. 3

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision*, 2014. 3

[23] Yiyo Kuo, Taho Yang, and Guan-Wei Huang, “The use of grey relational analysis in solving multiple attribute decision-making problems,” *Computers & Industrial Engineering*, vol. 55, no. 1, pp. 80–93, 2008. 3

[24] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang, “Universal style transfer via feature transforms,” *arXiv preprint arXiv:1705.08086*, 2017. 3, 4

[25] Weichen Fan, Jinghuan Chen, Jiabin Ma, Jun Hou, and Shuai Yi, “Styleflow for content-fixed image to image translation,” *arXiv e-prints*, pp. arXiv–2207, 2022. 3

[26] Zhizhong Wang, Zhanjie Zhang, Lei Zhao, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu, “Aesust: towards aesthetic-enhanced universal style transfer,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1095–1106. 3