Affective Image Filter: Reflecting Emotions from Text to Images

Shuchen Weng[#1,2,4] Peixuan Zhang[#3] Zheng Chang[3] Xinlong Wang[4] Si Li[*3] Boxin Shi[1,2]

[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

[2]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[3]School of Artificial Intelligence, Beijing University of Posts and Telecommunications

[4]Beijing Academy of Artificial Intelligence

{shuchenweng,shiboxin}@pku.edu.cn, {pxzhang,zhengchang98,lisi}@bupt.edu.cn, wangxinlong@baai.ac.cn

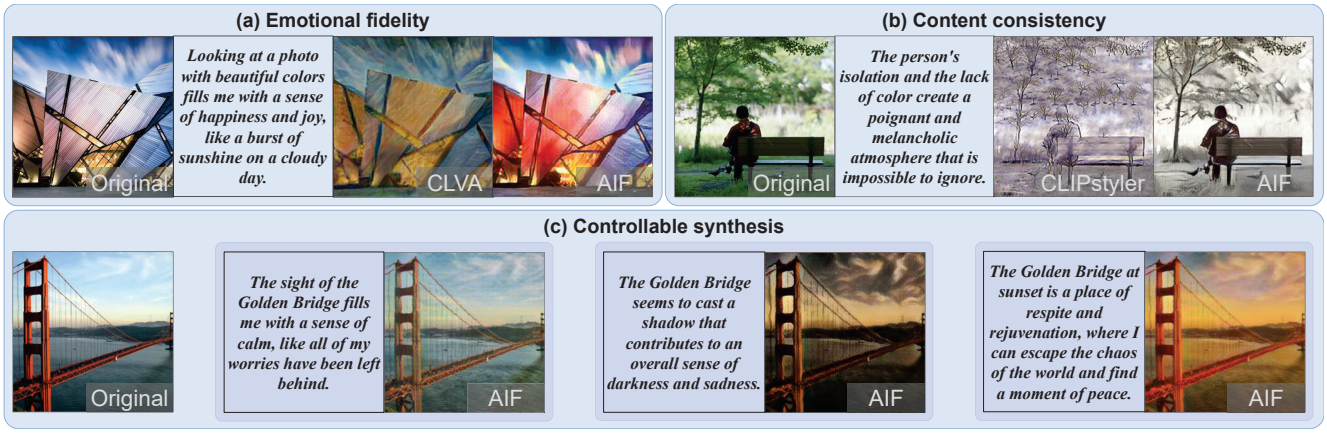

Figure 1: Illustration of the affective image filter (AIF) task. (a) Compared with CLVA [17], our AIF model acts as a bright and highly saturated image filter that tries to evoke specific emotional responses from human observers based on the given text. (b) Compared with CLIPstyler [26], our AIF model could maintain the visual elements, *e.g.*, humans and trees, while filling them with a distinct emotional tone. (c) Given an arbitrary image and different texts, our AIF model could synthesize a wide range of filtered images, without the need for individual fine-tuning or optimization of each input text [23, 26].

Abstract

Understanding the emotions in text and presenting them visually is a very challenging problem that requires a deep understanding of natural language and high-quality image synthesis simultaneously. In this work, we propose Affective Image Filter (AIF), a novel model that is able to understand the visually-abstract emotions from the text and reflect them to visually-concrete images with appropriate colors and textures. We build our model based on the multi-modal transformer architecture, which unifies both images and texts into tokens and encodes the emotional prior knowledge. Various loss functions are proposed to understand complex emotions and produce appropriate visualization. In addition, we collect and contribute a new dataset with abundant aesthetic images and emotional texts for training and evaluating the AIF model. We carefully design four quantitative metrics and conduct a user study to comprehensively evaluate the performance, which demonstrates our AIF model outperforms state-of-the-art methods and could evoke specific emotional responses from human observers.

1. Introduction

When people share their experiences about events and subjects on social networks (*e.g.*, Twitter), the text is a direct medium to express their opinions and establish emotional connections with other users [22, 43]. Since social media is a highly active platform with a vast amount of content being produced every day, influencers strive to personalize their content to evoke emotional responses from their followers.

It is well known that "a picture tells a thousand words". Images have powerful descriptive abilities, and they could also become affective stimuli that enable people from various backgrounds to understand emotional intention [66]. This motivates us to think that, given written texts that reflect personal thoughts and feelings, how we can reflect visually-abstract emotions from user-provided texts to

[#] Equal contributions. * Corresponding author.

visually-concrete images to further enhance the visual appeal of their social media posts.

In this work, we propose *Affective Image Filter* (AIF), a novel task that enables users to create unique and emotionally compelling images that "stand out from the crowd". The desired properties of a well-qualified AIF algorithm are outlined in Fig. 1, which demonstrates AIF's advantages if the following three objectives are met: *(i)* Emotional fidelity – The AIF model should accurately understand emotions from the text and reflect them in an arbitrary image provided by the user (Fig. 1 (a)). *(ii)* Content consistency – Since the AIF model acts as an image filter, it is required to preserve the overall structure and visual content of the content image provided by the user (Fig. 1 (b)). *(iii)* Controllable synthesis – Complementing to the above objectives *(i)* and *(ii)*, the AIF model should be capable of synthesizing results using a variety of emotional texts (Fig. 1 (c)).

To achieve these three objectives, we build the AIF model with the multi-modal transformer architecture which unifies both images as well as texts into tokens and encodes the prior knowledge of emotional words by utilizing the valence, arousal, and dominance (VAD) dictionary [40]. This prior knowledge assists our AIF model to have an in-depth understanding of the inherent properties behind the emotional words. Considering that low-level features (*i.e.*, colors and textures) of the visual content could well represent the evoked emotion [37, 65], we train the AIF model to learn the aesthetic style representations from famous paintings. The sentiment metric loss is designed to learn relationships between emotions; anchor-based sentiment loss and emotional distribution loss are designed to learn high-dimensional emotional cues; and other visualization losses are adopted to produce aesthetically pleasing images with appropriate colors and textures.

For training and evaluating the AIF model, we collect and contribute a new dataset with abundant images and corresponding text descriptions, where each text description could be categorized into one of Mikel's eight emotions[1] [38]. We further provide multiple quantitative metrics for evaluating whether the AIF model could achieve emotion-specific concrete visualizations of visually-abstract emotions in user-provided content images.

Our contribution could be summarized as follows:

- For the first time, we propose the AIF task to reflect visually-abstract emotions from text to images provided by the user and further provide metrics for comprehensive evaluation of performance.
- We introduce the prior knowledge of visual emotion analysis to develop the AIF model with transformer architecture and design novel losses to comprehensively visualize ambiguous and subjective emotions.
- An AIF dataset has been newly collected and processed. It includes numerous aesthetic images along with multiple emotional text descriptions associated with the closest emotional category.

[1]Mikel's eight emotions are: amusement, contentment, awe, excitement, fear, sadness, disgust, and anger.

2. Related Works

To achieve the desired properties of a well-qualified AIF solution, related work from three aspects need to be discussed: *(i)* **visual emotion analysis**, which plays an important role in producing corresponding concrete visualization with extracted emotional features and measuring whether the synthesized results could evoke specific emotional responses from human observers; *(ii)* **style transfer**, which aims to create aesthetically pleasing images based on a reference of colors and textures, while maintaining the visual content of user-provided content images, thereby sharing similar goals with the AIF task; and *(iii)* **vision transformer**, which has been demonstrated effective in modeling global token-to-token relationship between multi-modal inputs, making it a promising approach for interacting with user-provided images and texts in the AIF task.

2.1. Visual emotion analysis

Computer vision is increasingly focused on understanding emotion in context for more than two decades [27]. Before the advent of deep learning models, researchers developed a variety of handcrafted features for analyzing affective images [3, 37, 50, 61, 64, 65], which are typically vulnerable and difficult to generalize to out-of-distribution scenarios. This situation has been improved since neural networks are used to adaptively predict emotions [56, 62, 63], where models could extract multi-grained emotional features [45] or focus on local image regions [59]. On the basis of previous works [38, 55, 57], we could design novel constraints to learning the inherent semantics of emotions and makes synthesized results favorable by human observers.

2.2. Style transfer

Early works adopt the optimization-based method [19] or design end-to-end models [21, 29] to achieve style transfer for one specific style. To improve the efficiency of style transfer, researchers explore the approach to train multiple styles in one model [7, 16, 33], and further propose the first arbitrary style transfer model [20]. Using self-attention mechanisms to build long-range dependencies between regions [36, 41, 53] has received considerable attention in numerous studies devoted to improving the performance of arbitrary style transfer models [2, 9, 31, 48]. Following them, StyTr^2 [11] further adopts the transformer architecture to extract and maintain the global information of in-

put images. Recently, researchers have attempted to replace the reference image with semantic textures of text to provide more flexible and user-friendly style guidance [17, 26]. Compared to AIF algorithm, which is required to accurately understand *visually-abstract emotions* from the text, image style transfer methods aim to create images based on a reference of *visually-concrete colors and textures*.

2.3. Vision transformer

Transformer [49] has gained significant attention and popularity in recent years, leading to the development of unique feature fusion mechanisms for cross-modality tasks, *e.g.*, text-to-image generation [14, 44], visual grounding [10, 42], and referring segmentation [13, 60]. Meanwhile, researchers have explored the use of pure vision transformer models for a wide range of vision applications to achieve better performance, *e.g.* image classification [15], object detection [4, 68], and semantic segmentation [47, 67]. Great efforts have also been made to adapt vision transformer models to low-level vision problems, *e.g.*, inpainting [30, 35], super-resolution [8, 32], and colorization [5, 6, 52]. For better concrete visualization of visually-abstract emotions and an in-depth understanding of the inherent properties of emotional words, we adopt the transformer architecture to interact with cross-modal inputs.

3. AIF Dataset

To train a well-qualified AIF model, we collect and process a large-scale dataset that includes abundant aesthetic images with diverse colors and textures, corresponding text descriptions associated with the closest emotional category in the Mikel's wheel [38]. Although recently proposed ArtEmis [1] and ArtEmis v2 [39] datasets that aim to connect vision, language, and affection have similar goals to us, we find they are not inherently appropriate for the AIF task due to the following reasons: *(i)* A number of given text descriptions focus on clarifying the visual content of the corresponding image rather than providing detailed emotional descriptions of the observed individuals (Fig. 2 (a)); *(ii)* numerous text descriptions contain an excessive number of color and texture words that are useful in visualization, distracting the models from an in-depth understanding of the emotional words (Fig. 2 (b)). To prepare qualified data for training and evaluating, we manually select samples to create a new dataset tailored for the AIF task.

We begin by merging all the samples from ArtEmis [1] and ArtEmis v2 [39], which include anchor images that provide a fixed point of visually-concrete reference for visualizing colors and textures from visually-abstract emotions, as well as corresponding multiple emotional text descriptions. After that, we manually discard all unqualified text descriptions and verify that the remaining anchor images have an average of four to five corresponding text descriptions (Fig. 2 (c)). As emotions have ambiguity and subjectivity, we further categorize each text description into Mikel's eight emotions [38], which are used to measure the emotional distribution of each anchor image and establish inherent emotional relationships between them. As a result, the AIF dataset has emotional text descriptions with 16.3K amusement, 82.9K contentment, 43.1K awe, 22.0K excitement, 53.2K fear, 71.3K sadness, 26.9K disgust, and 9.5K anger samples. Since the visual content of images is a crucial feature for vision emotion analysis [37, 65], we additionally provide a similarity list for each image based on the VGG similarity calculation. Finally, we split the dataset into 69.6K anchor images and 292.9K emotional text descriptions for training, and 7.7K anchor images and 32.5K emotional text descriptions for evaluating.

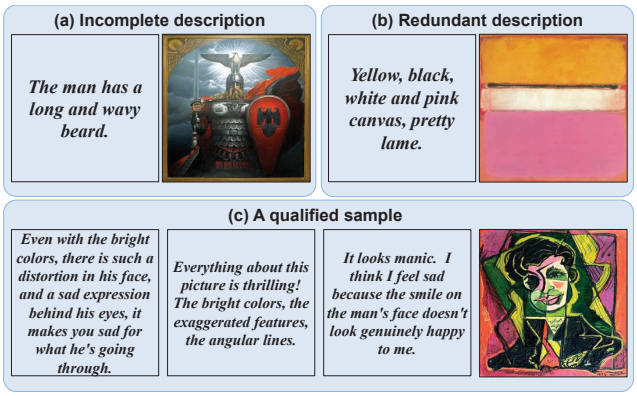

Figure 2: Examples of unqualified and qualified samples for the AIF task, where each sample includes text description(s) and a corresponding anchor image. (a) An unqualified sample with incomplete description, which focuses on clarifying the appearance of the human while ignoring its solemn feeling. (b) Another unqualified sample with redundant description, which provides too many color words for visualizations while distracting the models from understanding the emotional word "pretty". (c) A qualified sample includes balanced description of diverse subjective emotions expressed by human observers in response to the same anchor image.

4. AIF Model

In this section, we first present the overview of the AIF model, including data sampling strategy and the loaded data that provide different supervisory signals (Sec. 4.1). Next, we present the AIF transformer and elaborate on the detailed designs of modules (Sec. 4.2). After that, we show the approach to understand visually-abstract emotions (Sec. 4.3) and create concrete visualization of emotions (Sec. 4.4), followed by details of training settings (Sec. 4.5).

4.1. Overview

We sample text descriptions in the following steps: *(i)* We randomly select a number of text descriptions from the

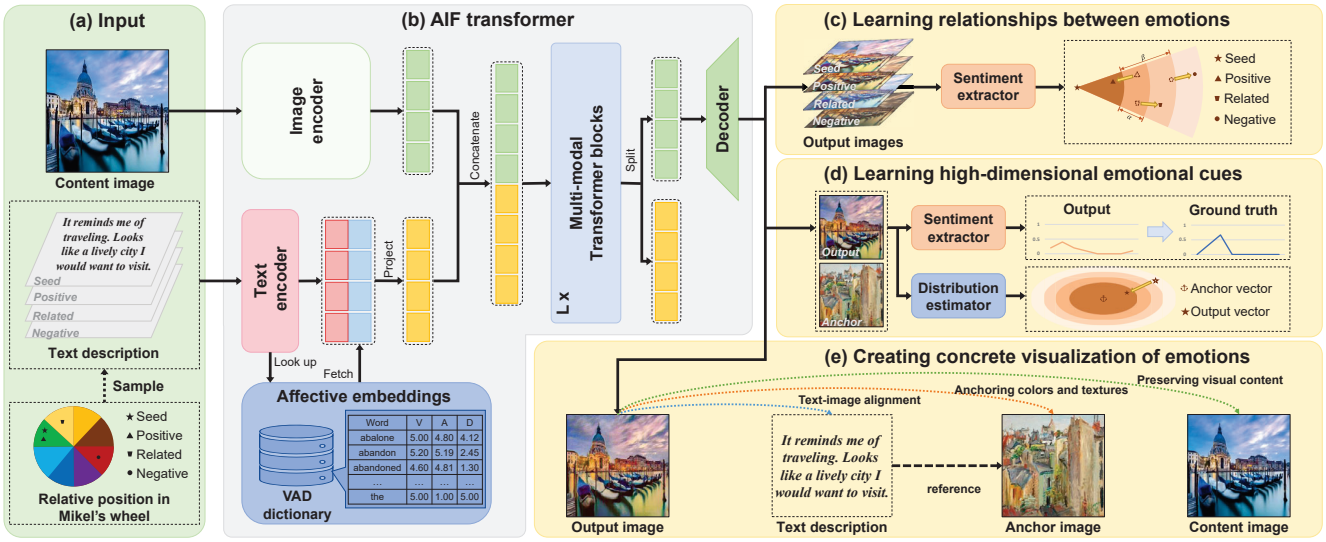

Figure 3: Overview of our proposed AIF model: (a) According to the relative position of Mikel's wheel [38], we sample combinations of text descriptions from the AIF dataset to reflect corresponding emotions to user-provided content images (Sec. 4.1). (b) We build the AIF model with transformer architecture that separately encodes user-provided content images and texts with inherent emotional properties into shared latent space and interacts them with each multi-modal transformer block (Sec. 4.2). (c) We define the distance between emotions, and design the sentiment metric loss to learn relationships between emotions (Sec. 4.3.1). (d) We design anchoring sentiment loss and emotional distribution loss to learn high-dimensional emotional cues (Sec. 4.3.2). (e) We apply various visualization losses so that the AIF model could align synthesized images with user-provided texts, visualize appropriate colors and textures, and preserve original visual content (Sec. 4.4).

AIF dataset as seed text descriptions. *(ii)* We categorize text descriptions based on their location in the Mikel's wheel [38], where closer regions have more similar valence or arousal. Text descriptions in the same region as the seed text descriptions are considered positive samples, those in adjacent regions are considered related samples, and those in opposite regions are considered negative samples. *(iii)* We refer to the similarity lists provided by the AIF dataset to select a positive, a related, and a negative text description for each seed text description. These selections and seed text descriptions collectively form texts in a data batch. As a result, we could present text descriptions in each batch as $[T^{\text{sed}}, T^{\text{pos}}, T^{\text{rel}}, T^{\text{neg}}]_i$ where $i \in \{1, \dots, \frac{1}{4}N_{\text{batch}}\}$. This assists our designed AIF transformer, which captures long-range dependencies between tokens in Eqs. (1-4), to learn relationships between emotions in Eqs. (5-6).

We further load corresponding anchor images for each text description, enabling our AIF model to use them as references for synthesizing images that evoke specific emotional responses from human observers, as presented in Eq. (7). Since emotions of images have ambiguity and subjectivity, we further calculate emotional distribution by normalizing corresponding emotional categories of text descriptions for each sampled anchor image, which assists the AIF model to reflect emotions to images more accurately, as shown in Eq. (8). Considering that anchor images also provide the reference for visualizing colors and textures, we further present the approach of creating concrete visualization of emotions under their guidance in Eqs. (9-12).

During training, we randomly select images from the MS-COCO dataset [34] as the content images. Note that only content images and corresponding text descriptions are fed to the AIF transformer, as shown in Fig. 3 (a); anchor images and corresponding emotional distributions are used as training guidance, which are not required during evaluating and inference.

4.2. AIF transformer

Instead of adopting an iterative optimization process, we build our AIF model based on the end-to-end training multi-modal transformer architecture to synthesize on-the-fly results. As the framework illustrated in Fig. 3 (b), our AIF transformer could be divided into the following four parts.

Image encoder. We split content images $I^{\text{cnt}} \in \mathbb{R}^{3\times H\times W}$ into patches $I^{\text{pat}} = [I_1^{\text{pat}}, \dots, I_N^{\text{pat}}] \in \mathbb{R}^{N\times P^2\times 3}$, where (P, P) is the patch resolution and $N = HW/P^2$. Following StyTr^2 [11], we use a transformer-based image encoder to capture long-range dependencies of image patches and output image embedding sequence $T^{\text{img}} = [T_1^{\text{img}}, \dots, T_N^{\text{img}}] \in \mathbb{R}^{N\times C_{\text{img}}}$, where C_{img} is the channel number.

Text encoder. We use the pre-trained BERT [12] to encode texts into word embeddings, and further fetch affective embeddings in VAD dictionary [40] to reveal the inherent emotional properties (*i.e.*, valence, arousal, and dominance) of

each word. For words missing in the VAD dictionary, we manually assign neural values. After that, these affective embeddings are concatenated with word embeddings as text tokens $T^{\text{tex}} = [T_1^{\text{tex}}, \ldots, T_M^{\text{tex}}] \in \mathbb{R}^{M \times C_{\text{tex}}}$, where M is the number of words and C_{tex} is the channel number.

Multi-modal transformer. We project image tokens and text tokens into shared latent space with MLP layers (f^{img} and f^{tex}, respectively). After that, we introduce the modal-type embedding vectors T_0^{typ} and T_1^{typ} to distinguish token modalities following ViLT [24], which are separately added to corresponding latent codes as:

$$\hat{T}_i^{\text{img}} = f^{\text{img}}(T_i^{\text{img}}) + T_0^{\text{typ}}, \ \hat{T}_j^{\text{tex}} = f^{\text{tex}}(T_j^{\text{tex}}) + T_1^{\text{typ}}, \quad (1)$$

where $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, M\}$. We denote the initial input of the multi-modal transformer as:

$$Z_0 = [\hat{T}_1^{\text{img}}, \ldots, \hat{T}_N^{\text{img}}, \hat{T}_1^{\text{tex}}, \ldots, \hat{T}_M^{\text{tex}}] \in \mathbb{R}^{(N+M) \times C_0}, \quad (2)$$

where C_0 is the channel number. The multi-modal transformer consists of L standard transformer block [15], and each block includes a multi-headed self-attention (MSA) layer, an MLP layer, and two residual connections. We formulate the process of each block as:

$$[\bar{Z}_i] = \text{MSA}(\text{LN}([Z_{i-1}])) + [Z_{i-1}], \quad i \in \{1, \ldots, L\} \quad (3)$$

$$[Z_i] = \text{MLP}(\text{LN}([\bar{Z}_i])) + [\bar{Z}_i], \quad i \in \{1, \ldots, L\} \quad (4)$$

where LN means the LayerNorm.

Image decoder. Following SETR [67], we build our image decoder as a three-layer CNN, which could alleviate the grid artifacts caused by the patch partition of images. Finally, we obtain the synthesized results $I^{\text{out}} \in \mathbb{R}^{3 \times H \times W}$.

4.3. Understanding visually-abstract emotions

By learning from the guidance in the AIF dataset, a well-qualified AIF model should understand visually-abstract emotions in user-provided texts and synthesize aesthetically pleasing images that evoke specific emotional responses from human observers. To achieve this goal, we train the model to learn the relationships between emotions and high-dimensional emotional cues.

4.3.1 Learning relationships between emotions

Based on the Mikel's wheel [38], we could define the distance between emotions, and further learn their relationships with our designed sentiment metric loss, as shown in Fig. 3 (c). Specifically, we first build the sentiment extractor following Yang *et al.* [57], which composes of a VGG network [46] that extracts the multi-level feature of synthesized images, a convolutional layer that projects features and multiple Gram matrices [18] that calculate the correlation between each pair of projected features. Next, we define the sentiment vector by concatenating elements as:

$$V = \text{Concat}_{i \in \{1, \ldots, N_{\text{gram}}\} \cup j \in \{1, \ldots, N_{\text{lev}}\}}(\Phi_{i,j}), \quad (5)$$

where $\Phi_{i,j}$ means the i-th upper triangular elements in the Gram matrix at j-th feature level, N_{gram} is the number of elements, and N_{lev} is the number of levels. Therefore, given combinations of texts $[T^{\text{sed}}, T^{\text{pos}}, T^{\text{rel}}, T^{\text{neg}}]_i$ (details in Sec. 4.1), we could extract corresponding sentiment vectors of synthesized images as $[V^{\text{sed}}, V^{\text{pos}}, V^{\text{rel}}, V^{\text{neg}}]_i$ with the sentiment extractor. According to the Mikel's wheel [38], we define the distance between emotions as the minimum number of steps from an emotion region to another, denoted as F_{dis}, so that the distance between sentiment vectors could be formulated as $\text{F}_{\text{sw}}(V_i, V_j) = \frac{\|V_i - V_j\|^2}{\text{F}_{\text{dis}}(V_i, V_j)}$. As a result, the AIF model learns relationships between emotions in a metric learning manner as:

$$\begin{aligned}\mathcal{L}_{\text{sm}} = &\max(\text{F}_{\text{sw}}(V^{\text{sed}}, V^{\text{pos}}) - \text{F}_{\text{sw}}(V^{\text{sed}}, V^{\text{rel}}) + \alpha), 0) \\ &+ \max(\text{F}_{\text{sw}}(V^{\text{sed}}, V^{\text{rel}}) - \text{F}_{\text{sw}}(V^{\text{sed}}, V^{\text{neg}}) + \beta), 0), \quad (6)\end{aligned}$$

where $\alpha = 0.02$ and $\beta = 0.01$ are hyper-parameters that control margins between sentiment vectors.

In practice, we employ anchor images to pre-train the sentiment extractor using the sentiment metric loss. Once the sentiment extractor is well-trained, its parameters are frozen. We then extract sentiment vectors of synthesized images with it, and optimize the distance between the sentiment vectors by reapplying the sentiment metric loss.

4.3.2 Learning high-dimensional emotional cues

As shown in Fig. 3 (d), we design the anchor-based sentiment loss and emotional distribution loss for learning high-dimensional emotional cues. The anchor-based sentiment loss utilizes anchor images as the reference to facilitate synthesized images in evoking specific emotional responses from human observers. Specifically, we extract sentiment vectors of each synthesized image and the corresponding anchor image with the well-trained sentiment extractor, denoted as V^{out} and V^{acr}, and require synthesized vectors to be close to corresponding anchor vectors as:

$$\mathcal{L}_{\text{as}} = \|V^{\text{out}} - V^{\text{acr}}\|_2. \quad (7)$$

Furthermore, considering that emotions have ambiguity and subjectivity (*e.g.*, Fig. 2 (c)), which causes each image could stimulate a range of emotional reactions to even a single person, we further design the emotional distribution loss to estimate the distribution of emotions, instead of categorizing each image into one dominant emotion. Specifically, we pre-train a distribution estimator φ, and use the Kullback-Leibler (KL) [25] loss to measure the information loss caused by the inconsistency between the estimated distribution and the ground truth as:

$$\mathcal{L}_{\text{ed}} = \sum_{i=1}^{N_{\text{cat}}} d_i \ln \frac{d_i}{\varphi(I^{\text{out}})_i}, \quad (8)$$

where N_{cat} is the number of emotional categories, $\varphi(I^{\text{out}})_i$ and d_i are value of i-th category of the estimated distribution and the ground truth, respectively. This assists the AIF model to reflect emotions from text to images accurately.

4.4. Creating concrete visualization of emotions

As shown in Fig. 3 (e), the visually-concrete images produced by the AIF model should adhere to a series of strict constraints. As the solution, we adopt the following visualization losses to meet these requirements:

GAN loss. A multi-level conditional-unconditional discriminator is designed to align synthesized images with user-provided texts, as well as to discriminate whether synthesized images are aesthetically pleasing, written as:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} &= \log D(I^{\text{acr}}) + \log\big(1 - D(G(I^{\text{pat}}, T^{\text{tex}}))\big) \\ &+ \log D(I^{\text{acr}}, T^{\text{tex}}) + \log\big(1 - D(G(I^{\text{pat}}, T^{\text{tex}}), T^{\text{tex}})\big),\end{aligned} \tag{9}$$

where the discriminator D consists of a stack of convolutional layers that extract image features and fully connected layers that project text tokens, and the generator G is our AIF model to synthesize image I^{out}. I^{pat} and T^{tex} are image patches and initial text tokens, respectively (details in Sec. 4.2). I^{acr} is the corresponding anchor images.

Style loss. Following style transfer models [11, 36, 41], we use the anchor image as the reference of colors and textures. With the pre-trained VGG network [46] to extract multi-level features of synthesized images and anchor images, style loss is adopted to narrow the style difference between extracted features as:

$$\mathcal{L}_{\text{s}} = \sum_i \|\mu(\phi_i^{\text{out}}) - \mu(\phi_i^{\text{acr}})\|_2 + \|\sigma(\phi_i^{\text{out}}) - \sigma(\phi_i^{\text{acr}})\|_2, \tag{10}$$

where μ and σ are the mean and variance functions, respectively. ϕ_i^{out} and ϕ_i^{acr} are extracted features of synthesized images and anchor images at i-th level, respectively.

Identity loss. We further utilize the identity loss, where we feed anchor images as content images and corresponding texts into the AIF model to synthesize identity images. As such, inputs and expected synthesized results have no gap in colors, textures, and visual contents, which encourages the model to learn richer and more accurate representations from input images. We require synthesized results to be consistent with original content images as:

$$\mathcal{L}_{\text{id}} = \|I^{\text{idt}} - I^{\text{acr}}\|_2 + \gamma \sum_i \|\phi_i^{\text{idt}} - \phi_i^{\text{acr}}\|_2, \tag{11}$$

where $\gamma = 0.01$ is a hyper-parameter, I^{idt} is identity images, and ϕ_i^{idt} is the corresponding extracted features.

Content loss. We present the content loss to preserve the visual content of user-provided content images, which narrows the squared error of extracted features at each level as:

$$\mathcal{L}_{\text{c}} = \sum_i \|\phi_i^{\text{out}} - \phi_i^{\text{cnt}}\|_2, \tag{12}$$

where ϕ_i^{cnt} is the features of content images at i-th level.

4.5. Training details

We train our AIF model with full objective losses by solving a minimax optimization problem as:

$$\begin{aligned}\max_D \min_G &\lambda_{\text{sm}}\mathcal{L}_{\text{sm}} + \lambda_{\text{as}}\mathcal{L}_{\text{as}} + \lambda_{\text{ed}}\mathcal{L}_{\text{ed}} \\ &+\lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{c}}\mathcal{L}_{\text{c}},\end{aligned} \tag{13}$$

where we set hyper-parameters as $\lambda_{\text{sm}} = 30$, $\lambda_{\text{as}} = 600$, $\lambda_{\text{ed}} = 140$, $\lambda_{\text{GAN}} = 3$, $\lambda_{\text{s}} = 0.3$, $\lambda_{\text{id}} = 2$, and $\lambda_{\text{c}} = 5$ based on experiments using a held-out validation set, which are not sensitive to variations in a certain range.

All experiments are conducted on 4 NVIDIA TITAN RTX GPUs and trained for 80K iterations with batch size 24 for 30 hours. We use the Adam optimizer to minimize losses with the warm-up adjustment strategy [54], and set the learning rate as 5×10^{-4}.

5. Experiment

5.1. Quantitative evaluation metrics

To comprehensively evaluate the performance of models for the AIF task, we adopt the following four quantitative metrics: *(i)* Following CLVA [17], we use **Structural Similarity Index Measure (SSIM)** [51] to measure whether synthesized images have similar visual content with content images. *(ii)* Following StyTr^2 [11], we calculate the **Style Difference (SD)** between synthesized images and anchor images to measure whether synthesized images have appropriate colors and textures. *(iii)* We extract sentiment vectors of synthesized images and anchor images with the pre-trained sentiment extractor [57], and calculate the Euclidean distance between them to measure whether synthesized images could evoke specific emotional responses, denoted as **Sentiment Gap (SG)**. *(iv)* With the pre-trained distribution estimator, we calculate the **Accuracy (Acc)** as Yang *et al.* [58] to measure whether synthesized images accurately reflect visually-abstract emotions.

5.2. Comparison with state-of-the-art methods

As our model is the first trial for the AIF task, we conduct comparison experiments with related image editing methods (*i.e.*, ManiGAN [28] and DiffusionCLIP [23]) and style transfer methods (*i.e.*, CLIPstyler [26] and CLVA [17]).

Qualitative comparisons. We show visual quality comparisons with the methods above in Fig. 4. Among these methods, ManiGAN [28] distorts the visual content (the first row, a sharp boundary appears in the sky); DiffusionCLIP [23] fails to preserve image semantics (the second row, the city street turns into the building); results of CLIPstyler [26] are overly stylized (the third row, the train presents an unnatural color tone). CLVA [17] tends to synthesize colorless results regardless of emotional cues (the fourth row, the night

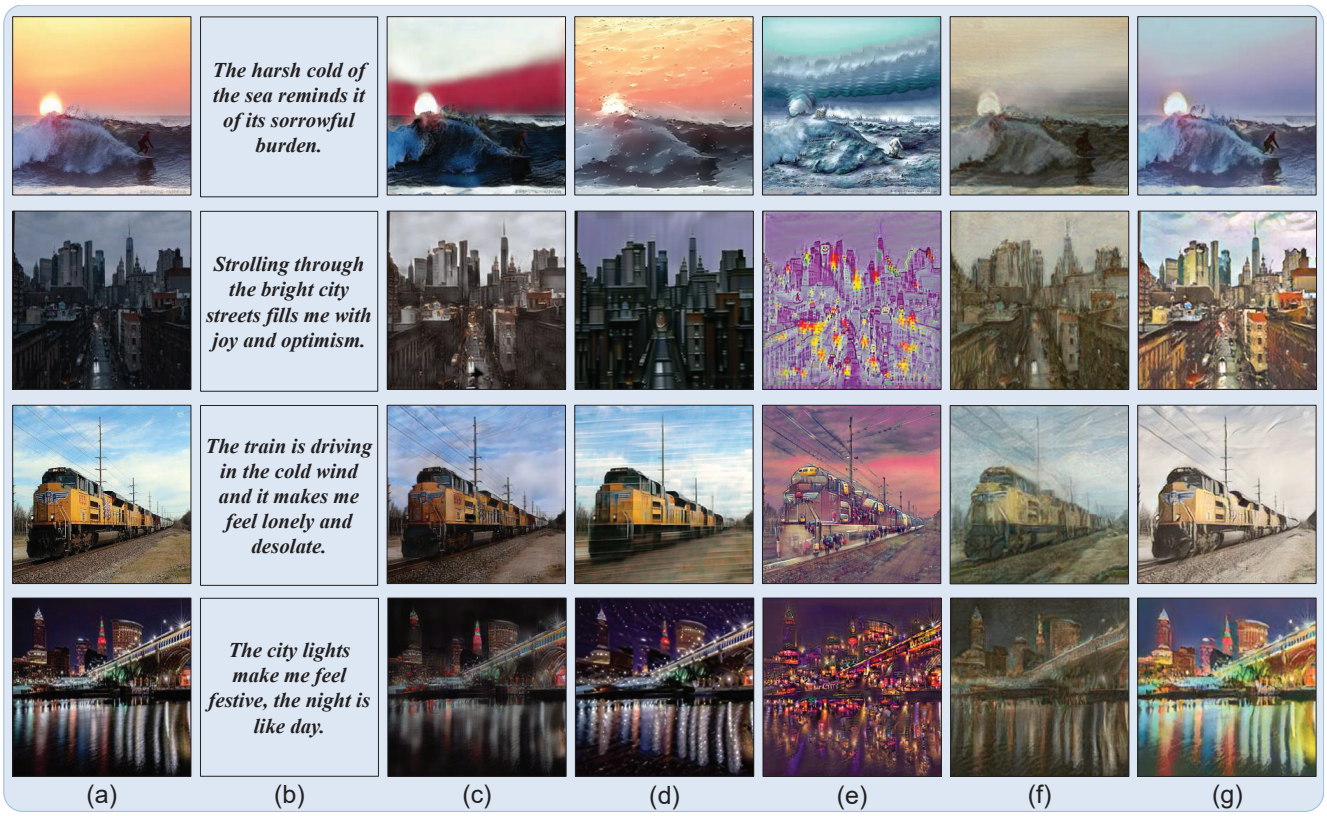

Figure 4: Qualitative comparison results with state-of-the-art methods. (a) User-provided content images. (b) Texts that reflect thoughts and feelings. (c) ManiGAN [28]. (d) DiffusioinCLIP [23]. (e) CLIPstyler [26]. (f) CLVA [17]. (g) Our results.

Table 1: Quantitative experiment results of comparison and ablation. ↑ (↓) means higher (lower) is better. Best performances are highlighted in **bold**.

Comparison with state-of-the-art methods				
Method	SSIM (%) ↑	SD ↓	SG (‰) ↓	Acc (%) ↑
ManiGAN	50.72	7.8913	1.6589	27.77
DiffusionCLIP	53.05	10.6151	1.7095	24.59
CLIPstyler	52.49	10.4493	1.5676	26.40
CLVA	50.30	6.3715	1.6707	25.64
Ours	**56.15**	**5.4147**	**1.3881**	**29.96**
Ablation study				
W/o VAD	54.75	5.8416	1.4284	29.76
W/o SE	55.62	5.5876	1.4799	29.56
W/o ED	53.75	5.7672	1.3915	26.48
W/o GAN	56.07	5.5148	1.3911	29.15

scene seems being covered with grey dust). Our method better understands emotions from the text and reflects them to images with appropriate colors and textures.

Quantitative comparisons. We show quantitative comparisons in Tab. 1, where the highest scores on all metrics demonstrate that our method outperforms compared state-of-the-art methods. Specifically, our method well preserves the original visual content (SSIM), synthesizes appropriate colors and textures (SD), evokes specific emotional responses from human observers (SG), and reflects visually-abstract emotions more accurately (Acc).

Table 2: User study results. Our method outperforms other approaches with the highest score.

Method	ManiGAN	DiffusionCLIP	CLIPstyler	CLVA	Ours
Preference	9.08	13.24	10.32	12.68	**54.68**

User study. In addition to qualitative and quantitative comparisons, we further conduct a user study experiment to find out whether images synthesized by our model are preferred by human observers over compared state-of-the-art methods. Each sample shown to participant consists of a content image, an emotional text, and five synthesized images. Participants are then asked to select the synthesized image that best matches the emotional text. The experiment is published on Amazon Mechanical Turk (AMT), where 100 samples from the testing set of the AIF dataset are randomly selected, and experiment results are polled independently by 25 volunteers. As a result, our model achieves the highest preference score, which demonstrates the subjective advantages of our approach. The preference scores

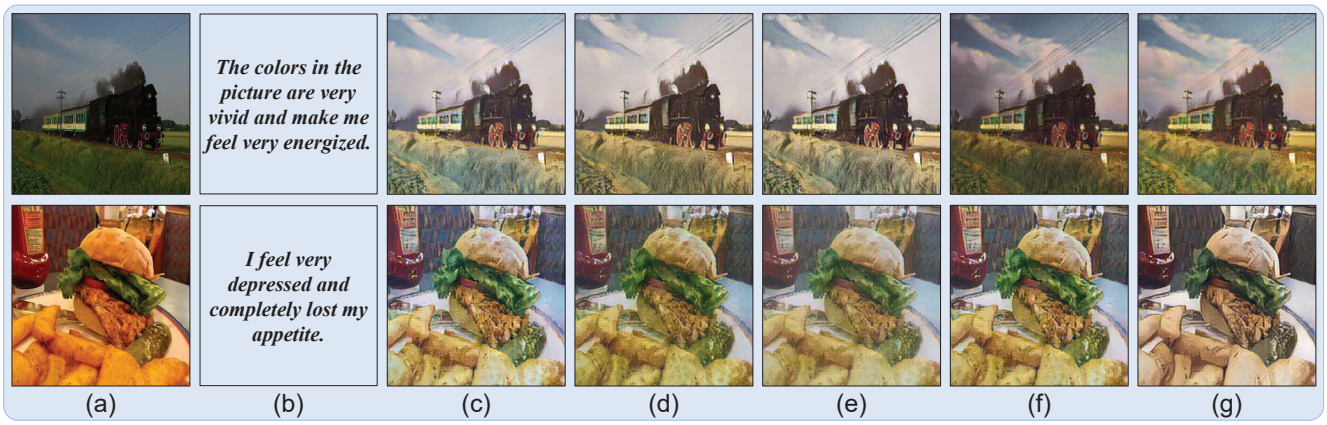

Figure 5: Ablation study results with different variants of the proposed method. (a) User-provided content images. (b) Texts that reflect thoughts and feelings. (c) W/o VAD. (d) W/o SE. (e) W/o ED. (f) W/o GAN. (g) Our results.

are shown in Tab. 2.

5.3. Ablation study

We discard various modules and create four baselines to study the impact of our proposed modules and designed losses. The evaluation scores and synthesized images of the ablation study are shown in Tab. 1 and Fig. 5, respectively.

W/o VAD. We remove the VAD dictionary in the text encoder, which causes the AIF model cannot obtain the prior knowledge of inherent emotional properties. This increases the difficulty in synthesizing appropriate colors and textures (higher SD score. As shown in the first row of Fig. 5, the filter appears washed out due to low saturation and high brightness, instead of being vivid).

W/o SE. We disable the sentiment extractor along with all related designs (*e.g.*, sentiment metric loss and anchor-based sentiment loss). This leads to a wider sentiment gap between synthesized results and anchor images (lower SG score. As shown in the second row of Fig. 5, the food appears moldy and looks uncomfortable with the green filter, but it does not evoke a sense of depression).

W/o ED. We discard the emotional distribution estimator and the emotional distribution loss, which prevent the AIF model from learning the ambiguous and subjective distribution of emotions. As a result, the AIF model fails to accurately reflect emotions from text to images (reduced Acc score. As shown in the second row of Fig. 5, this filter is unremarkable and fails to evoke clear emotional response).

W/o GAN. We remove the discriminator that aligns synthesized images with user-provided texts and pushes the model to produce aesthetically pleasing images. This degrades the overall performance and results in text-image misalignment (decreased overall scores. As shown in the first row of Fig. 5, the dark tone makes people feel weighed down and exhausted, instead of being energized).

6. Conclusion

We for the first time propose the AIF task to reflect emotions from text to images. To capture long-range dependencies between cross-modal inputs, we build our AIF model based on the multi-modal transformer architecture. Various novel losses are designed for better understanding complex emotions and creating appropriate visualization. We further collect and process a dataset tailored for the AIF task. In addition, we carefully design four quantitative metrics and conduct a user study to demonstrate that our AIF model outperforms related state-of-the-art methods and could synthesize aesthetically pleasing results that evoke specific emotional responses from human observers.

Overall, the proposed AIF task and model present a promising avenue for future research in the field of cross-modal emotion understanding and image synthesis. Our dataset and metrics can serve as an evaluation protocol, for assessing the future generative foundation models' performance in emotion understanding and visualization.

Limitation. The performance of the AIF model may be limited by the suitability of the visual content provided by the user. Specifically, if the user-provided image is not well-suited to express the intended emotion of the text, *e.g.*, attempting to reflect sadness to a picture of a child's innocent smiling face, the AIF model may have difficulty in producing convincing results.

7. Acknowledgements

This work is supported by the National Key R&D Program of China under Grant No. 2021ZD0109800, the National Natural Science Foundation of China under Grand No. 62136001, 62088102, Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications, and program for Youth Innovative Research Team of BUPT No. 2023QNTD02.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. ArtEmis: Affective language for visual art. In *CVPR*, 2021. 3

[2] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. ArtFlow: Unbiased image style transfer via reversible neural flows. In *CVPR*, 2021. 2

[3] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[5] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. L-CoDer: Language-based colorization with color-object decoupling transformer. In *ECCV*, 2022. 3

[6] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-CoIns: Language-based colorization with instance awareness. In *CVPR*, 2023. 3

[7] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. StyleBank: An explicit representation for neural image style transfer. In *CVPR*, 2017. 2

[8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 3

[9] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. In *NIPS*, 2021. 2

[10] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-end visual grounding with transformers. In *ICCV*, 2021. 3

[11] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr2: Image style transfer with transformers. In *CVPR*, 2022. 2, 4, 6

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 4

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 3

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. CogView: Mastering text-to-image generation via transformers. In *NIPS*, 2021. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 5

[16] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017. 2

[17] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *ECCV*, 2022. 1, 3, 6, 7

[18] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In *NIPS*, 2015. 5

[19] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[22] Jaap Kamps, Maarten Marx, Robert J Mokken, and Maarten de Rijke. Using wordnet to measure semantic orientations of adjectives. In *LREC*, 2004. 1

[23] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022. 1, 6, 7

[24] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 5

[25] Solomon Kullback and Richard A Leibler. On information and sufficiency. *AoMS*, 1951. 5

[26] Gihyun Kwon and Jong Chul Ye. CLIPstyler: Image style transfer with a single text condition. In *CVPR*, 2022. 1, 3, 6, 7

[27] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1997. 2

[28] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. ManiGAN: Text-guided image manipulation. In *CVPR*, 2020. 6, 7

[29] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016. 2

[30] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-aware transformer for large hole image inpainting. In *CVPR*, 2022. 3

[31] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NIPS*, 2017. 2

[32] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV*, 2021. 3

[33] Minxuan Lin, Fan Tang, Weiming Dong, Xiao Li, Changsheng Xu, and Chongyang Ma. Distribution aligned multimodal and multi-domain image stylization. *TMM*, 2021. 2

[34] 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 *ECCV*, 2014. 4

[35] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. FuseFormer: Fusing fine-grained information in transformers for video inpainting. In *ICCV*, 2021. 3

[36] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *ICCV*, 2021. 2, 6

[37] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010. 2, 3

[38] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *BRM*, 2005. 2, 3, 4, 5

[39] Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *CVPR*, 2022. 3

[40] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *ACL*, 2018. 2, 4

[41] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *CVPR*, 2019. 2, 6

[42] Mengxue Qu, Yu Wu, Wu Liu, Qiqi Gong, Xiaodan Liang, Olga Russakovsky, Yao Zhao, and Yunchao Wei. SiRi: A simple selective retraining mechanism for transformer-based visual grounding. In *ECCV*, 2022. 3

[43] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *ACL*, 2020. 1

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 3

[45] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *NPL*, 2020. 2

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 5, 6

[47] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 3

[48] Jan Svoboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *CVPR*, 2020. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[50] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACM MM*, 2022. 2

[51] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 2004. 6

[52] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. CT^2: Colorization transformer via color tokens. In *ECCV*, 2022. 3

[53] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *ICCV*, 2021. 2

[54] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *ICML*, 2020. 6

[55] Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. A circular-structured representation for visual emotion distribution learning. In *CVPR*, 2021. 2

[56] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018. 2

[57] Jufeng Yang, Dongyu She, Yu-Kun Lai, and Ming-Hsuan Yang. Retrieving and classifying affective images via deep metric learning. In *AAAI*, 2018. 2, 5, 6

[58] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 2017. 6

[59] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Visual sentiment prediction based on automatic discovery of affective regions. *TMM*, 2018. 2

[60] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 3

[61] Xingxu Yao, Dongyu She, Haiwei Zhang, Jufeng Yang, Ming-Ming Cheng, and Liang Wang. Adaptive deep metric learning for affective image retrieval and classification. *TMM*, 2020. 2

[62] Quanzeng You, Hailin Jin, and Jiebo Luo. Visual sentiment analysis by attending on local image regions. In *AAAI*, 2017. 2

[63] Wei Zhang, Xuanyu He, and Weizhi Lu. Exploring discriminative representations for image emotion recognition with cnns. *TMM*, 2019. 2

[64] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 2014. 2

[65] Sicheng Zhao, Hongxun Yao, You Yang, and Yanhao Zhang. Affective image retrieval via multi-graph learning. In *ACM MM*, 2014. 2, 3

[66] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *TPAMI*, 2021. 1

[67] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3, 5

[68] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2020. 3