

Article

Exploring Emotional Stimuli Detection in Artworks: A Benchmark Dataset and Baselines Evaluation [†]

Tianwei Chen ^{1,*} , Noa Garcia ¹ , Liangzhi Li ²  and Yuta Nakashima ¹ ¹ Intelligence and Sensing Lab, Osaka University, Suita, Osaka 565-0871, Japan; noagarcia@ids.osaka-u.ac.jp (N.G.); n-yuta@ids.osaka-u.ac.jp (Y.N.)² Computer Science Department, Qufu Normal University, Qufu 273165, China; liliangzhi@ieee.org

* Correspondence: chentw@is.ids.osaka-u.ac.jp

[†] This paper is an extended version of our paper published in ICMR 2024, Phuket, Thailand, 10–14 June 2024.

Abstract: We introduce an emotional stimuli detection task that targets extracting emotional regions that evoke people's emotions (i.e., emotional stimuli) in artworks. This task offers new challenges to the community because of the diversity of artwork styles and the subjectivity of emotions, which can be a suitable testbed for benchmarking the capability of the current neural networks to deal with human emotion. For this task, we construct a dataset called APOLO for quantifying emotional stimuli detection performance in artworks by crowd-sourcing pixel-level annotation of emotional stimuli. APOLO contains 6781 emotional stimuli in 4718 artworks for validation and testing. We also evaluate eight baseline methods, including a dedicated one, to show the difficulties of the task and the limitations of the current techniques through qualitative and quantitative experiments.

Keywords: emotional stimuli detection; artwork analysis



Citation: Chen, T.; Garcia, N.; Li, L.; Nakashima, Y. Exploring Emotional Stimuli Detection in Artworks: A Benchmark Dataset and Baselines Evaluation. *J. Imaging* **2024**, *10*, 136. <https://doi.org/10.3390/jimaging10060136>

Academic Editors: Eva Cetinic and Sinem Aslan

Received: 30 April 2024

Revised: 26 May 2024

Accepted: 28 May 2024

Published: 4 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Analyzing artworks in machine learning is a challenging task. Compared to photographs, artworks do not only depict real-world concepts, such as humans, animals, and natural scenes, but also represent humane contents, such as feelings, attitudes, and faiths. The richness in the representations and the diversity of styles make artworks the ideal testbed to study new challenges related to human emotion understanding in machine learning.

In recent years, many efforts have been undertaken in the field of artwork analysis [1–10], aiming to improve models' understanding of artworks and further extend models' capacity to support digital humanities, with tasks such as attribute identification [1,5–8], object detection [3,4], or artwork understanding through language [2,9,10]. Thanks to these studies, recent models have developed a reliable capacity to understand *objective* contents (e.g., objects, attributes, and descriptions) from the artworks. However, only a few studies [11,12] are focusing on a more *subjective* and personal analysis, such as the relationship between artwork and emotions.

ArtEmis [11] and its extension ArtEmis V2.0 [12] are two datasets collected for studying the relationship between artworks and emotions. The main focus is on the generation of emotional captions that can accurately capture the emotional influence of an artwork. However, a more in-depth analysis to uncover why and how emotions are evoked from the artworks is still not explored. In other words, ArtEmis and ArtEmis 2 show that models can generate emotional captions, but it is still unknown how the emotions are evoked from those artworks.

Artworks can easily elicit people's emotions, yet this elicitation process is complex and underexplored [13–16]. The appraisal theory toward artworks and emotions [13] says the emotion-evoking process is related to the viewer's analysis process; the emotions are evoked during the viewer's analysis process through the whole artwork. Thus, different analyses may lead to different emotions. For example, given the artwork in Figure 1,

people may feel different emotions when the analysis concentrates on different visual concepts in the context of the artwork. If a viewer focuses on the distorted style of the person, a feeling of amusement may be evoked, while the bear-like brown figures may be linked with fear. Learning such processes could make models acquire knowledge about how human emotions are evoked and may improve models' capacity to utilize emotional stimuli. Such merits could be helpful for tasks related to emotions (e.g., visual emotion recognition [17–21]) and tasks potentially involving emotion analysis (e.g., image generation [22,23]).

According to these observations, we propose a new task for emotional stimuli detection in artworks, in which a model is required to detect emotional stimuli from a given artwork, as shown in Figure 1. The task, which explores a machine's capacity to understand emotions and artworks, has two major challenges: First, different from photorealistic images, *artworks are painted with a certain style*. For example, in Western art, Realism is one of the styles that may look more like a real photo, while Impressionism typically shows prominent brush strokes. Different styles lead to very different appearances of the same object. This style variation makes it harder to learn visual content (e.g., objects) from artworks than photos [1,4,24]. Second, *emotions are subjective*. Different people may have different emotions evoked by the same artwork [11,12]. This subjectivity makes the task unique, as an artwork can have multiple emotional stimuli for different emotions.

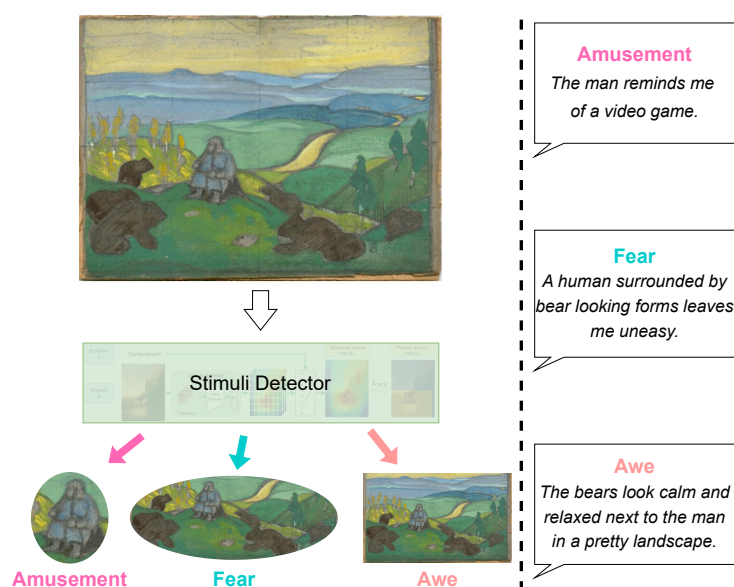


Figure 1. A model detects emotional regions that evoke people's emotions (namely, emotional stimuli) from the given artwork. The utterances on the right side may be used as hints to spot emotional stimuli.

For this task, we construct a benchmark dataset, coined APOLO (**Artwork Provoked emOtion EvaLuatiOn**), to evaluate models in both qualitative and quantitative ways. We build APOLO on top of the ArtEmis dataset [11], which, for each artwork, includes emotion labels annotated by multiple annotators, and utterances (sentences) that explain why emotions are provoked. To further understand the stimuli that provoke emotions, APOLO includes pixel-level emotional stimuli annotations on the images of the test set. As a result, we collect 6781 emotional stimuli for 4718 artworks and eight emotions. Our exhaustive control quality checks ensure the samples are balanced and reliable. To the best of our knowledge, this is the first dataset that offers pixel-level annotations of emotional stimuli in artworks.

Additionally, we explore multiple models for emotional stimuli detection, borrowed from related tasks, including object detection, referring expression, and saliency map detection. We also introduce a dedicated weakly supervised model as a baseline, which predicts

emotional stimuli regions for each emotion without using region-level annotations for training. Our comprehensive experiments on APOLO show that the evaluated models can detect emotional stimuli even when not trained with region annotations. However, the emotional stimuli detection task is still challenging and has plenty of room for improvement. In addition, we explore how a text-to-image generative model, Stable Diffusion [22], handles emotions in the input prompts, observing that it fails to connect the emotional words in the input with the emotional stimuli in the generated images. We hope our work will help overcome this limitation in the future. Our data and code are available at <https://github.com/Tianwei3989/apolo>.

2. Related Work

Visual emotion analysis Given an input image, visual emotion analysis aims to recognize emotions, analyze the emotional stimuli, and apply the recognized emotions to real-world applications (e.g., psychological health [25,26] and opinion mining [27,28]) to improve the ability of emotional intelligence [29]. Most of the recent studies [17–21] use emotional stimuli to improve emotion recognition, but only a few efforts have been made to analyze how well the models detect such stimuli. To the best of our knowledge, only two datasets, the EmotionROI [30] and EMOd [31], provide pixel-level annotations for evaluating emotional stimuli detection. However, they are both relatively small, offering 1980 and 1019 labeled images, respectively, consisting of social media images from the Internet.

Data scarcity is one of the main challenges in emotional stimuli detection. To overcome this problem, we propose two solutions: (1) transferring models from related tasks and (2) designing a weakly supervised learning model that does not require costly pixel-level annotations for training. For evaluation, we collect a dataset with emotional stimuli annotations.

Artwork analysis Much effort has been dedicated to solving art-related problems with machine learning techniques, including style identification [1,32], object detection [1,3,4], instance-level recognition [33], or artwork description [2,9,34]. Concerning emotion analysis, some datasets [1,11,12,35], including ArtEmis, contain labels with the emotion (e.g., *amusement* and *fear*) that each artwork evokes. Nevertheless, the same artwork can evoke multiple emotions according to different regions of the image, a fact that has been unexplored in current datasets. APOLO introduces a new challenge by investigating the connection between artworks and emotion at the pixel level.

3. Emotional Stimuli Detection

Our task aims to explore how a model can find the cues of the emotion elicitation process from the artwork, i.e., the emotional stimuli. In general, we explore two separate scenarios: (1) emotional stimuli detection *without reference* (i.e., utterances) and (2) emotional stimuli detection *with reference*. Ideally, a model should find emotional stimuli without reference, like humans can. However, such models are rare since only a few studies are aimed at emotional stimuli detection. We thus also explore whether recent multimodal models can detect emotional stimuli by using the references.

Formally, let $a, e \in \mathcal{E}$, and u denote an artwork, its emotion label, and the utterance, which can be a set of sentences, in ArtEmis, where $\mathcal{E}$ is the set of the emotions. $\mathcal{D}_t$ denotes the training set of ArtEmis [11], where $\mathcal{D}_t$ contains triples (a, e, u) . $\mathcal{D}_v$ and $\mathcal{D}_e$ denote the validation and test sets of APOLO, where both contain triples (a, e, u) . As we presume that an artwork can evoke potentially *any* emotion depending on where the viewer focuses their attention, the emotional stimuli detection task can be formulated as a segmentation task given artwork a and emotion $e \in \mathcal{E}$, in which a model f_e predicts segment s that evoke emotion e as

$$\hat{s} = f_e(a), \quad (1)$$

where $\hat{s}$ is the predicted segment.

This task can be extremely challenging as no cue is provided for specifying the regions that are involved given emotion e . We thus formulate a variant with reference by u , in which u is given to a model as an auxiliary cue for emotional stimuli detection, i.e.,

$$\hat{s} = f_e(a, u). \quad (2)$$

In both scenarios (emotional stimuli detection with and without reference), we can use $\mathcal{D}_t$ for training a model, but ground-truth segments are not available. $\mathcal{D}_v$ and $\mathcal{D}_e$ are solely used for validation and testing.

4. APOLO Dataset Curation

APOLO is a benchmark dataset for evaluating, both quantitatively and qualitatively, emotional stimuli detection in artworks. We utilize the test samples in ArtEmis [11], with 39,850 explanatory utterances and emotional responses related to 8003 artworks from WikiArt (<https://www.wikiart.org/> (accessed on 30 May 2024)).

ArtEmis is annotated with nine emotions: *amusement*, *anger*, *awe*, *contentment*, *disgust*, *excitement*, *fear*, *sadness*, and *something else*. As shown in Figure 2, the utterances are explanations of why a certain emotion is evoked by the artwork, which is usually related to its emotional stimuli.



Figure 2. Some samples in our dataset. The words and regions in green are the chosen noun phrases and the annotated emotional stimuli, respectively. If one artwork-emotion pair contains multiple utterances, the corresponding regions are then combined. We annotate regions for eight emotions from ArtEmis, except "something else".

We observe that the utterances potentially align with the viewers' analysis processes and are related to a certain emotion that is specified by the emotion label and evoked by the artwork. Furthermore, the utterances tend to describe certain regions, which leads to a certain emotion, in the artwork. These features may help models to learn how humans perceive emotion from or associate emotion with such regions.

Toward this end, we construct a pixel-level emotional stimuli dataset, APOLO, by asking 91 annotators in Amazon Mechanical Turk (<https://www.mturk.com/>) (accessed on

30 May 2024) to identify the visual concepts that involve the utterances and to annotate the visual concepts at the pixel level, as shown in Figures 2 and 3. We show its details in Table 1.

Table 1. Summary of datasets for emotional stimuli detection. The source column indicates whether the images are from the social internet or artworks, while the ME (multi-emotion) column indicates whether an image has annotations for multiple emotions.

	Samples	Images	Source	Emotions	ME
EmotionROI [30]	1980	1980	social	6	No
EMOd [31]	1019	1019	social	2 + 1	No
APOLO	6781	4178	artwork	8	Yes

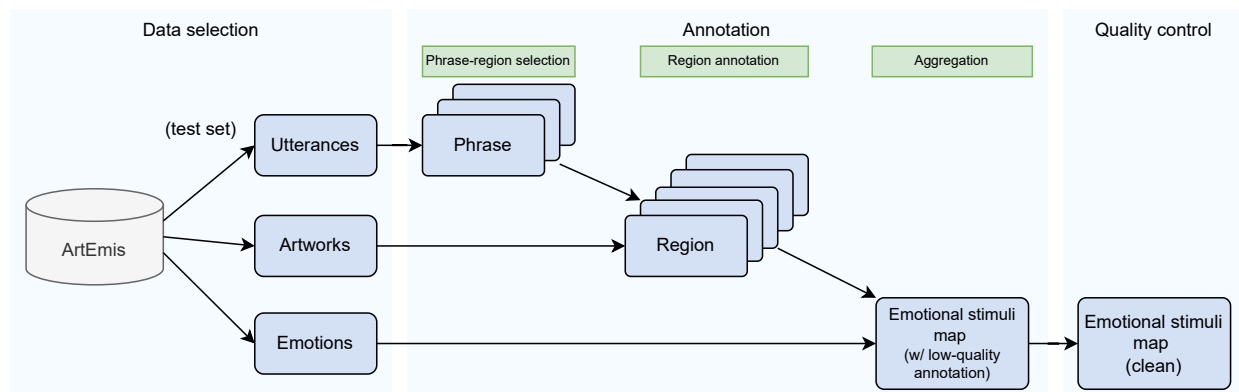


Figure 3. The workflow of APOLO dataset curation. In general, APOLO is extended from the ArtEmis dataset. We collect the annotation from ArtEmis’s test set and further annotate the pixel-level emotional stimuli map from the artworks.

We only collect validation and test sets by randomly sampling ArtEmis’s test set, considering (1) the evaluation could be applied to recent large models (e.g., CLIP [36]) that are hard to train and (2) the cost of pixel-level annotation. During the APOLO dataset curation, we always use random selection as our selection principle.

4.1. Data Selection

To curate our annotations from ArtEmis, we annotate paintings from the first eight emotions from ArtEmis’s nine emotions, i.e., the emotions of *amusement*, *anger*, *awe*, *contentment*, *disgust*, *excitement*, *fear*, and *sadness*. We filter out samples with the *something else* label, as we found from the associated utterance that their interpretation of the emotion is not trivial and annotators may not capture clear ideas from them. For each of the other eight emotion labels, we randomly choose about 1200 artwork–utterance pairs from the ArtEmis test set, except for the emotion *anger*, which only contains 672 artwork–utterance pairs. Overall, we select 9599 samples.

4.2. Annotation Process

As the aim is to annotate emotional stimuli, which are regions that can evoke a certain emotion, we design an annotation process focused on identifying the regions that correspond to specific phrases in the utterances, as these phrases are strongly tied to the emotions. The general annotation process is shown in Figure 4, and it consists of three steps: (1) phrase-region selection, (2) region annotation, and (3) aggregation.

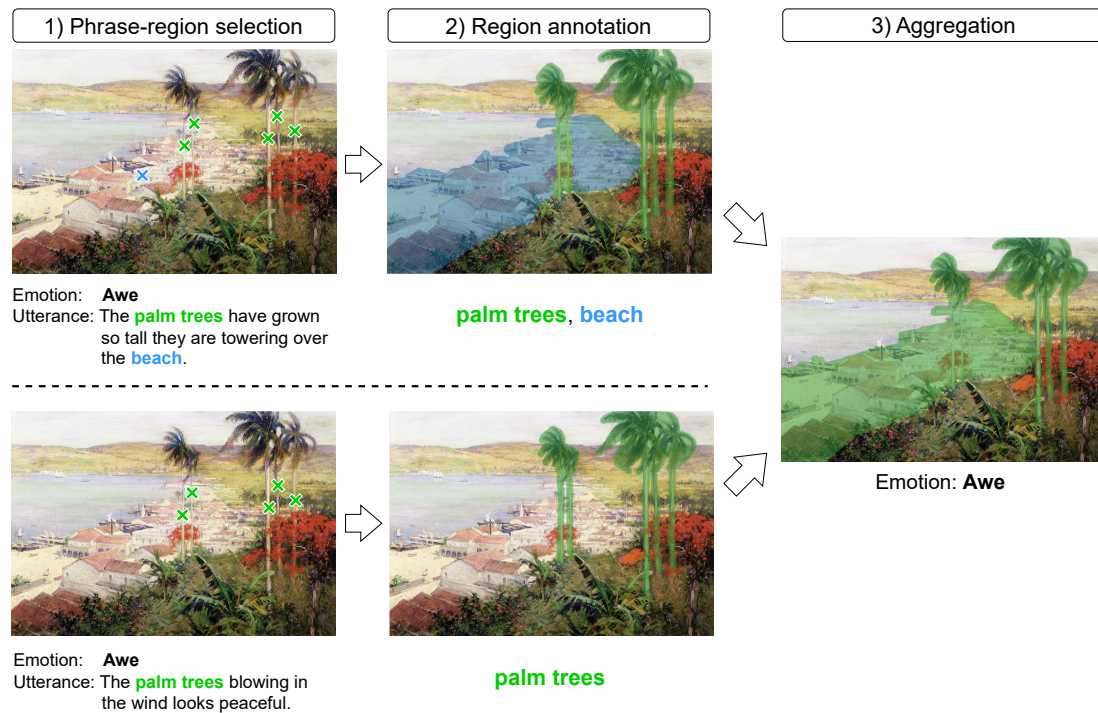


Figure 4. In our annotation process, workers should annotate the following three steps: (1) phrase-region selection, (2) region annotation, and (3) aggregation. The green and blue ×'s in the first step are the location of the regions. We randomly check the submissions at every step to ensure the annotation quality.

Phrase-region selection The annotation interface is shown in Figure 5. In the first step, we aim to gather the cues of the emotion elicitation process from the utterances, i.e., to collect phrases in the utterances that correspond to the emotional stimuli and their corresponding location on the artwork.

To identify such phrases, we show annotators a single utterance u together with an artwork a . Note that by design, ArtEmis utterances u explicitly describe the emotion generated by the artwork a . Then, we ask them to find all the noun phrases in u that explicitly mention visual concepts in a . We denote the set of identified phrases in u by W_u , where $w \in W_u$ is a phrase (e.g., the “trees”). Specifically, we provide annotators two additional options, the *whole artwork* and the *nothing to label* (the yellow and orange buttons as shown in Figure 5), since some utterances may only talk about the whole image or nothing related to the artwork. If there is at least one phrase in the utterances that corresponds to the emotional stimuli, the annotators are then asked to locate the region in the artwork by spotting at least one point that lies in the region of the visual concepts by clicking on the artwork in our annotation interface. The set of points for phrase w is denoted by P_w , where $p \in P_w$ is in $\mathbb{R}^2$. To ensure that all phrases in the utterance and visual concepts are found, and also to reduce the subjectivity of annotation, we ask for three annotators per triplet (a, e, u) and aggregate annotations by removing duplicates to form W_u and P_w for all $w \in W_u$. By this step, we collect two types of annotations: (1) the noun phrases that are related to both the artworks and the evoked emotions (the colored phrases in Figure 4) and (2) the locations of the region that the noun phrases (the green and blue ×'s in Figure 4).

Region annotation In this step, we aim to identify the regions related to the emotion elicitation process, i.e., to draw pixel-level annotations according to the utterances. We collect these annotations based on the locations in the previous step. We show a , u , w , and P_w to an annotator and ask them to draw on top of a all pixels that fall into the visual concepts identified by $w \in W_u$ and P_w , obtaining a segment s_w , which is a set of pixels. By

this step, we collect pixel-level annotations of the regions for each of the noun phrase (the colored regions in Figure 4).

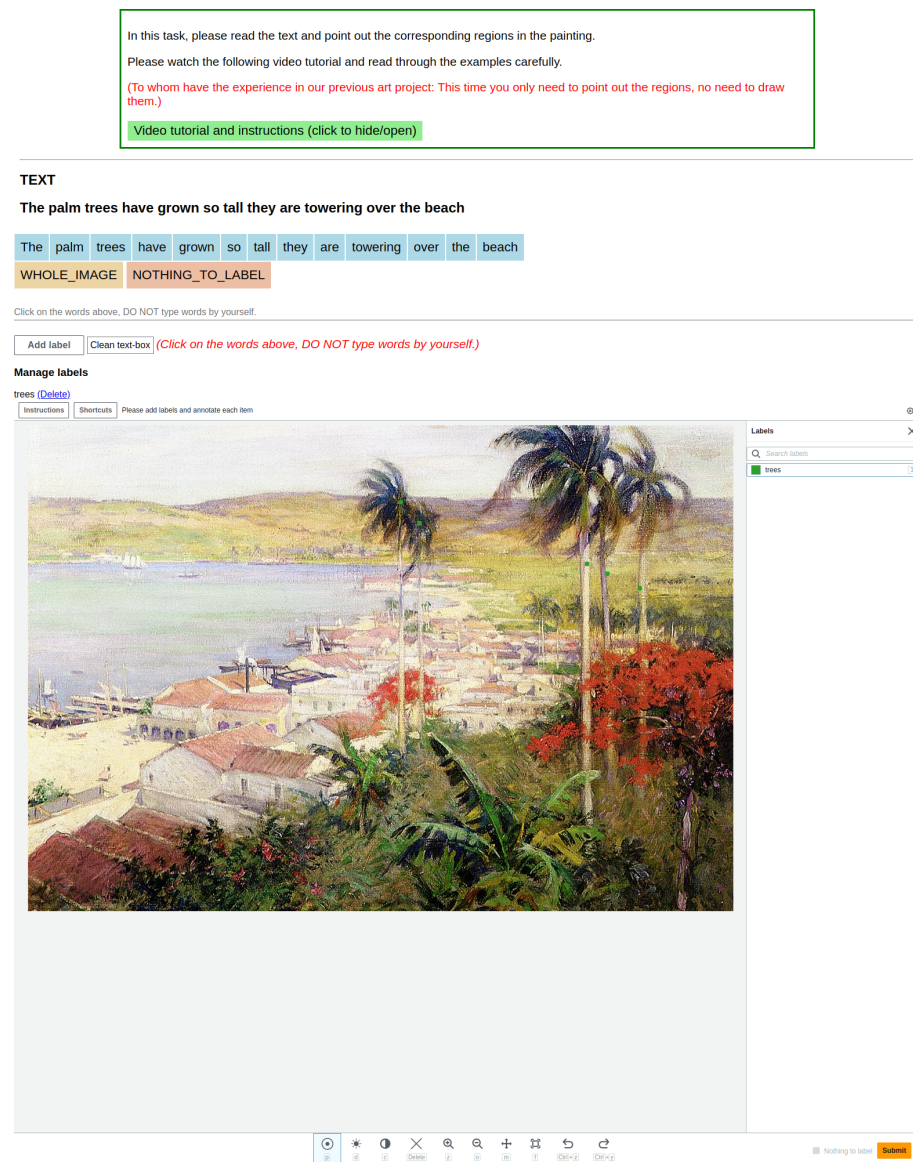


Figure 5. Annotation interface of phrase-region selection. On this interface, an annotator should first read the utterance and artwork and then point out the location of the region. The blue, yellow, and orange blocks in the “TEXT” section are the buttons for annotators to select.

Aggregation Next, we aggregate phrase-wise region annotations s_w belonging to the same a and emotion e . For all w that is associated with a and e , i.e., $w \in W = \{w \in W_{u'} | (a', e', u') \in \mathcal{D}, a = a', e = e'\}$, we obtain the aggregated emotional stimulus s by

$$s = \bigcup_{w \in W} s_w. \quad (3)$$

By this step, we finally collect the region annotations for each of the emotion. Some examples of a , u , and s are shown in Figure 2. As a result, we obtain 7512 emotional stimuli in 5160 artworks. The data structure is shown in Table 2.

Table 2. Data structure of our evaluation set.

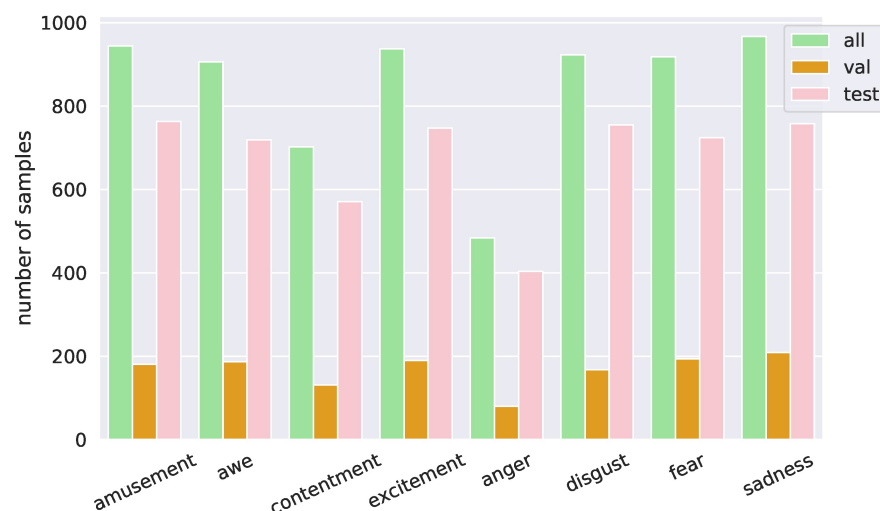
Painting	Emotion	Map_ID
a.y.-jackson_indian-home-1927	sadness	000000
aaron-siskind_new-york-24-1988	anger	000001
abdullah-suriosubroto_bamboo-forest	contentment	000002
abdullah-suriosubroto_mountain-view	excitement	000003
abraham-manievich_moscow-iii	excitement	000004
abraham-manievich_moscow-iii	sadness	000005
...	...	...

4.3. Quality Control

We apply quality controls both during and after the annotation process. During the annotation process, we randomly check 10% of the annotations in every round of submission and reject the *dishonest* ones (e.g., phrase $w \in W$ is wrong, region s_w is wrong, etc.). After the annotation process, we manually check all the annotations with special attention to the following three cases: (1) when the *whole artwork* is annotated as a region, (2) when the annotation is *low-quality* (e.g., only draw the contour) or wrong (e.g., draw wrong regions), and (3) when no region (denoted *void*) is annotated in the artwork. We found 1211 instances of *whole artwork*, 33 of *low-quality*, and 87 of *void* annotations. We remove all of them from our dataset. Finally, to ensure that the dataset is balanced and the *whole artwork* annotations are not over-represented, we randomly remove 600 *whole artwork* annotations to form our APOLO dataset.

4.4. Evaluation Dataset Analysis

APOLO consists of 6781 emotional stimuli for 4718 artworks. We split it into validation and test sets with approximately 20% and 80% of the samples, respectively. The artworks in the validation and the test sets are disjointed. Figure 6 shows the distribution of emotion label e in APOLO. We remark that seven out of eight emotions have more than 500 samples, while the number of *anger* samples is smaller due to the fewer samples in the original ArtEmis dataset. The distributions of the validation and test sets are similar to that of the entire dataset.

**Figure 6.** Emotion distribution of APOLO.

We also calculate the distribution of the ratio of pixels in s over a , i.e., $|s|/|a|$, where $|\cdot|$ gives the number of pixels in the region s or artwork a . Figure 7 shows the distribution. Many regions (46.94%) are small ($|s|/|a| \leq 0.375$), and less regions (24.01%) are large ($|s|/|a| > 0.625$). From this, our evaluation dataset tends to have regions that focus on

local concepts. The distributions of $|s|/|a|$ for the validation and test sets are also similar to the entire dataset. Similar to ArtEmis, one artwork may contain a varied number (from one to eight) of emotions, and one artwork–emotion pair may contain a varied number of utterances.

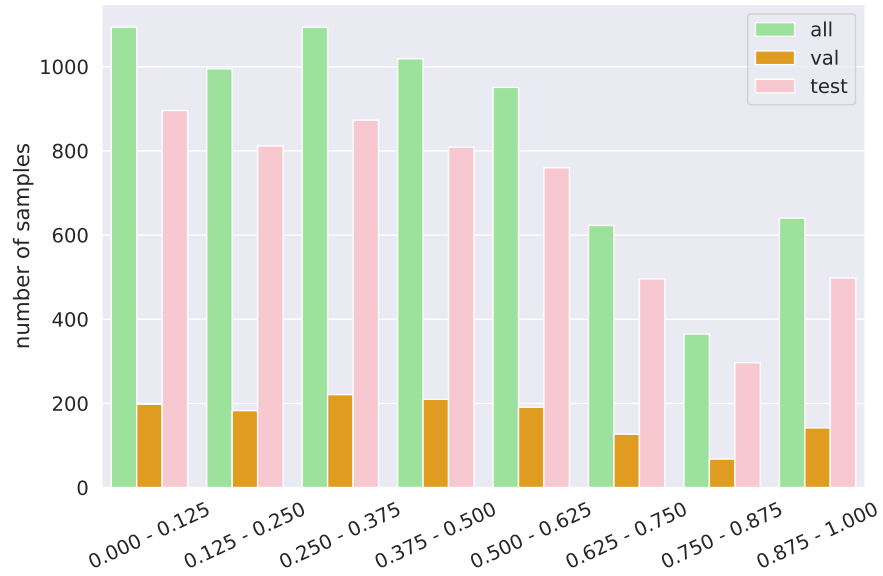


Figure 7. Stimuli occupation distribution of APOLO. The x-axis is the ratio of annotated pixels to the whole image, i.e., the occupation of the stimuli.

5. Baselines

To better comprehend the challenges of the emotional stimuli detection task, we propose and evaluate several baselines.

5.1. Baselines with Reference

In the with-reference variant, utterance u provides abundant information about what a model should look for, which reduces the task close to visual grounding, like refCOCO [37] and refCOCOg [38]. Our strategy is first to find regions relevant to u with utterance-region similarities and to weight the regions with the similarity to obtain a *emotional stimuli map* with pixel-level scores for e . This process is shown in Figure 8. Prediction $\hat{s}$ can be generated by thresholding the map.

We employ **VilBERT** [39] and **12-in-1** [40] as baselines, where 12-in-1 may have a variety of knowledge as it is trained over 12 vision-and-language tasks, while VilBERT is pre-trained on a large-scale dataset GCC [41]. To adapt to our task, VilBERT and 12-in-1 models are fine-tuned with refCOCO. These models give the probability of each region proposal given u , which can be interpreted as an utterance-region similarity score. **CLIP + VinVL** is a combination of CLIP [36] and VinVL [42]. CLIP [36] is renowned for its zero-shot capacity to solve vision-and-language tasks. We can first use VinVL to find region proposals and use CLIP to compute the utterance-region similarity with u .

Emotional stimuli map generation Let R denote the set of regions obtained from any of the above methods, and $\text{sim}(r, u)$ be the utterance-region similarity between $r \in R$ and u . We aggregate all regions in R to generate an *emotional stimuli map* M_u for u by

$$M_u = \sum_{r \in R} \text{sim}(r, u) m_r, \quad (4)$$

where m_r is a map that represents r by giving 1 if a pixel in m_r is in r and 0 otherwise. As an artwork a can be associated with multiple utterances for the same emotion, we aggregate all of them to obtain emotional stimuli map M_e for e as

$$M_e = \sum_{u \in U_{ae}} M_u, \quad (5)$$

where $U_{ae} = \{u' | (u', a', e') \in \mathcal{D}, a' = a, e' = e\}$. Thresholding is applied to M_e .

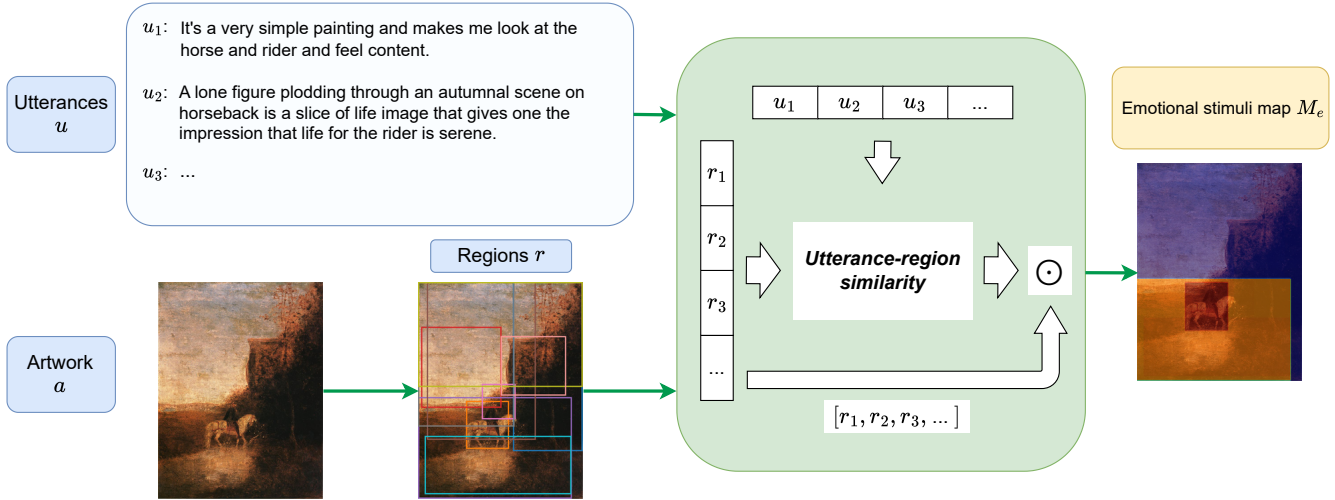


Figure 8. The overview of emotional stimuli map generation for baselines with reference.

5.2. Baselines without Reference

5.2.1. Object Detection

One naïve idea for the without-reference task is to spot salient regions in some senses and give the regions as emotional stimuli regardless of given emotion e . Object detection can give such regions [43]. We adopt the region proposal networks in **FasterRCNN** [44] and **VinVL** [42]. VinVL's region proposal network may offer better performance as it can additionally detect some attributes (e.g., *blue* and *calm*) that may exhibit stronger ties with some emotions. We aggregate proposals with top- K confidence to form $\hat{s}$ for any $e \in \mathcal{E}$ (i.e., $f_e(a) = f_{e'}(a)$ even for $e \neq e'$). To obtain segment prediction $\hat{s}$, we follow the same procedure as emotional stimuli map generation in the previous section, but we use $1/|r|$ in place of $\text{sim}(r, u)$ as this task does not allow us to use u , so we cannot compute $\text{sim}(r, u)$.

5.2.2. CASNet and CASNet II

CASNet [31] is a learning-based model for saliency detection, which generates a saliency map for a given image. The model is trained on a dataset called EMOD, which contains images that evoke some emotions and human fixations. With this dataset, CASNet learns to find regions that draw human attention. The work [31] showed, based on their analysis over EMOD, that humans tend to focus more on *emotional* objects than on *neutral* objects, where *emotional* and *neutral* objects are annotated by annotators. Therefore, CASNet also tends to focus on *emotional* objects. For our task, we apply thresholding to the saliency map to obtain $\hat{s}$. Again, prediction $\hat{s}$ is the same for all e . We also evaluate **CASNet II** [45], an extension of CASNet with atrous spatial pyramid pooling [46].

5.2.3. Weakly-Supervised Emotional Stimuli Detector

As the baselines for the without-reference task so far are not designed for this task and are ignorant of emotion label e , we design a dedicated model, abbreviated as WESD (**Weakly-supervised Emotional Stimuli Detection**), using utterances in ArtEmis [11] for weakly-supervised training.

An overview of WESD is shown in Figure 9. It first uses a visual encoder, such as ResNet variants [47], that gives patch-wise visual features. The visual features of respective patches of artwork a are then fed into a binary classifier for each e to predict if the patch

contains emotional stimuli for emotion e . Let v_i be a feature vector for patch $i \in K$, where K is the total number of patches in one artwork. Classifier g_e for emotion e predicts a score as

$$\hat{y}_{ei} = g_e(v_i) \in [0, 1]. \quad (6)$$

Specifically, g_e predict $\hat{y}_{ei}$ by the feature of both the certain patch and the whole artwork, as

$$g_e(v_i) = F_e \left(v_i + F_g \left(\frac{1}{K} \sum_k v_k \right) \right), \quad (7)$$

where $F_g(\cdot)$ is a fully-connected layer for embedding the whole artwork, and $F_e(\cdot)$ is a fully-connected layer to predict $\hat{y}_{ei}$. WESD contains multiple $F_e(\cdot)$'s and each of $F_e(\cdot)$ is related to a certain emotion e (e.g., *contentment*).

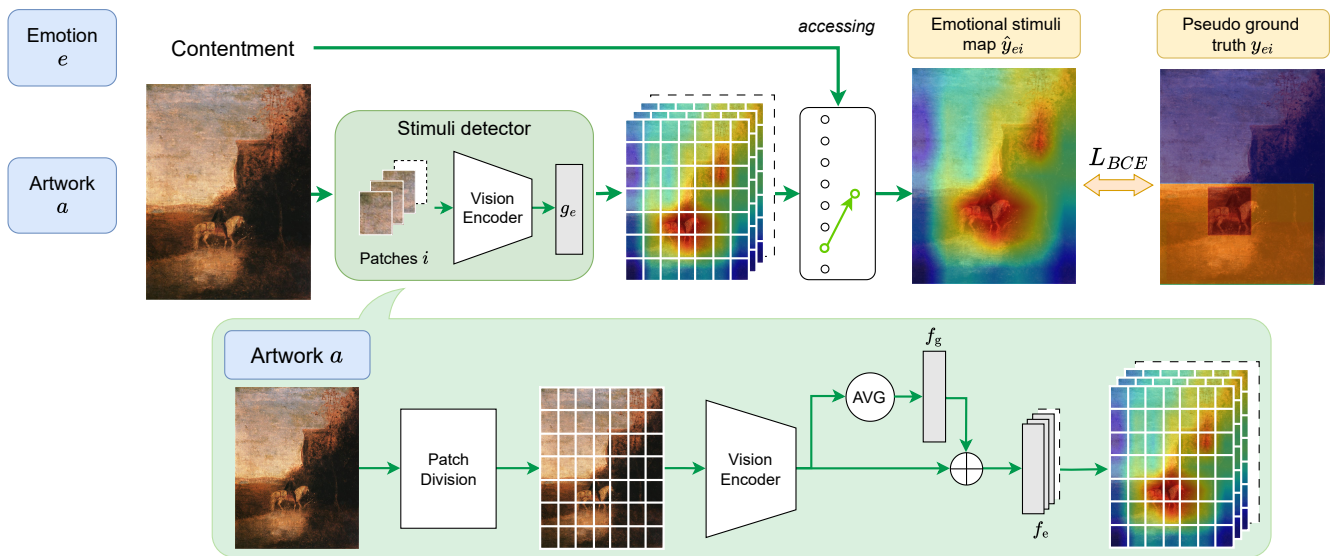


Figure 9. The overview of WESD. WESD predicts an emotional stimuli map for each emotion, and we access a certain map when the emotion (e.g., *contentment*) is given. For training, we use pseudo ground truth from CLIP + VinVL since APOLO does not have training data.

For training, ground-truth emotional stimulus s in APOLO can give direct supervision over y_{ei} ; however, APOLO is only for validation and testing. We instead use a pseudo ground truth. We utilize CLIP + VinVL for the with-reference task to generate an emotional stimuli map M_e , which can be derived from the ArtEmis training set. This means that the predictions based on utterances are used to distill the knowledge about the emotional stimuli into f_e for the without-reference (without-utterance) task. Emotion label e is only for identifying the map that has pseudo ground truth.

For this, we first divide M_e into the same patches as v_i 's and compute the mean within each patch to obtain a soft label y_{ei} . The binary cross-entropy loss $\text{BCE}(\cdot, \cdot)$ is used for training, i.e.,

$$L = \text{BCE}(\hat{y}_{ei}, y_{ei}). \quad (8)$$

For inference, WESD takes artwork a as the only input. The classifiers for all $e \in \mathcal{E}$ predict the score $\hat{y}_{ei}$, which is then summarized into $\hat{y}_e \in [0, 1]^{B_w \times B_h}$, where B_w and B_h are the numbers of patches in the horizontal and vertical axes, respectively. The map $\hat{y}_e$ is then resized to the same size as a to obtain predicted emotional stimuli map $\hat{Y}_e$. Predicted segment $\hat{s}$ can be obtained by thresholding over $\hat{Y}_e$.

6. Experiments

Metrics For evaluation, we borrow the ideas from previous works to employ *bounding box* [30] and *segmentation* [31] scenarios, where the former only requires to roughly locate emotional stimuli, while the latter requires their precise shapes. The *bounding box* [30] evaluation focuses on both stimuli and their background (e.g., the emotion of *awe* in Figure 1), as it assumes that emotions are evoked not only by the stimuli but also by the background. While the *segmentation* [31] evaluation focuses on the stimuli (e.g., the human and the bears in Figure 1) themselves, as it assumes that the stimuli are more important than other regions to evoke the certain emotions. We use both methods, in considering that both could be related to the emotion elicitation process. For both scenarios, we calculate the precision with intersection over union (IoU) threshold θ ($\text{Pr}@ \theta$), as in [37–40,48,49]. We evaluate models with $\text{Pr}@25$ and $\text{Pr}@50$.

For baselines that output bounding boxes (i.e., FasterRCNN and VinVL), we collectively treat them as a single region (though they can be disconnected) for evaluation in the segmentation scenario. In contrast, for baselines that give segments by thresholding, we generate a bounding box for each connected component for the bounding box scenario.

Implementation details Our baselines in most cases use the default setting in the original paper. As for CLIP, we use the ResNet-50 variant throughout our experiments. For WESD, we resize artworks to 224×224 pixels. Bi-linear interpolation is used to resize $\hat{y}_e$ to $\hat{Y}_e$. We train the model for 20 epochs with batch size 128, learning rate 2×10^{-4} , and decay 0.01. The model is optimized with AdamW [50]. For ViLBERT [39] and 12-in-1 [40], we follow the procedure in the respective papers to fine-tune the models on refCOCO [37]. All training processes were performed on a Quadro RTX 8000 GPU, which took about 20 h for WESD.

All baselines require a suitable threshold to obtain segment prediction $\hat{s}$. We use the APOLO validation set to find the best one on it with $\text{IoU}@50$ and apply it for evaluation.

Baseline variants For the baselines with reference, which take utterances as input, we can instead use the emotion label (e.g., *excitement*), so that the models can find regions that can be associated with (or that the models learned to associate with) the word.

In addition to the baselines in Section 5, we evaluate the case where the entire artwork is predicted as $\hat{s}$.

6.1. Quantitative Analysis

The scores of all baselines for both with-reference and without-reference tasks are summarized in Table 3. We list our findings as follows.

Artworks have something in common with natural images with respect to emotion. For the without-reference task, VinVL, CASNet, and WESD work well. CASNet is the best among these three models in terms of $\text{Pr}@25$ in both bounding box and segmentation scenarios. It also hits the second-best in $\text{Pr}@50$ of segmentation. Although marginal, the superiority of CASNet may imply that EMOd [31] used for training the model in a fully supervised manner has something in common with APOLO. This is intriguing as regions in images that seem to be in very different domains (i.e., natural images and paintings in various artistic styles) share some characteristics. This insight may elicit further exploration of the connection between natural images and paintings, like studying the types of paintings for which a model learned from EMOd works.

Emotional stimuli are highly correlated with objects and attributes. The scores of region proposals by both FasterRCNN and VinVL are still comparable to CASNet and WESD. For the metrics that require precise localization (i.e., $\text{Pr}@50$) and segmentation, the gap seems slightly larger. We would say that emotional stimuli highly coincide with some objects. This is reasonable because the utterances (e.g., in Figure 1) mention some objects. A comparison between FasterRCNN and VinVL suggests the correlation between VinVL attributes [42] and emotion, which again makes a lot of sense.

Table 3. Results of the evaluation on eight baseline models as well as a lower-bound baseline (i.e., the *Entire artwork*). From the left, *Task* tells that if a baseline model makes predictions with references (i.e., emotion tags or utterances). *Region proposal* shows if and which region proposal network is used by the certain baseline model. *Text input* shows the text input for generating emotional stimuli maps. *Multiple maps* is ✓ when the model can output multiple emotional stimuli maps for different emotions. For results in both *Bounding box* and *Segmentation*, we bold the best score and take an underline to the second-best result.

Task	Baseline	Region Proposal	Input Text	Multiple Map	Bounding Box		Segmentation	
					Pr@25	Pr@50	Pr@25	Pr@50
1	Entire artwork	-	-	-	82.37	63.81	68.61	37.70
2	FasterRCNN	FasterRCNN	-	-	84.03	66.40	74.43	43.67
3	w/o	VinVL	-	-	84.43	67.54	75.10	43.94
4	reference	CASNet	-	-	84.84	66.02	76.40	<u>44.15</u>
5	CASNet II	-	-	-	84.84	63.59	76.24	40.18
6	WESD	-	-	✓	84.30	<u>66.66</u>	75.89	44.97
7	VilBERT	FasterRCNN	emotion	✓	82.17	63.08	72.14	39.64
8	12-in-1	FasterRCNN	emotion	✓	72.51	50.71	63.90	31.87
9	w/	CLIP + VinVL	emotion	✓	81.97	63.00	71.29	40.05
10	reference	VilBERT	utterance	✓	84.10	65.41	75.26	42.18
11	12-in-1	FasterRCNN	utterance	✓	80.52	59.16	72.52	37.99
12	CLIP + VinVL	VinVL	utterance	✓	83.31	64.99	72.58	40.08

The domain of the utterances may be different from the text in typical vision-and-language tasks. Interestingly, the scores of the with-reference task are mostly lower than those of the without-reference task. This is counterintuitive as the utterances should give beneficial information to identify the emotional stimuli. One possible rationalization is the domain gap. The utterances in ArtEmis [11,12] come with subjective statements (e.g., “... bear looking forms leaves me uneasy.” in the second utterance in Figure 1), which is not likely in typical vision-and-language tasks. This observation can be supported by the fact that the use of the emotion labels, which is very different from typical text, as input to the vision-and-language models worsens the performance (lines 7–9 versus lines 10–12 in Table 3). Additionally, the worst scores of 12-in-1 also can also support this because the 12-in-1 model is fine-tuned to 12 vision-and-language tasks and may lose the generalization capability for unseen tasks.

WESD achieved better performance than CLIP + VinVL. Regardless of the worse performance of CLIP + VinVL, WESD hit a higher performance than it, although WESD is trained from CLIP + VinVL. A possible explanation is that, despite the lower performance of CLIP + VinVL for individual artworks, there are some characteristics shared in the dataset, and WESD may capture them through training.

6.2. Qualitative Analysis

Qualitative examples are shown in Figures 10 and 11 for the bounding box and segmentation scenarios, respectively. Figure 10 shows baselines with the top-two scores, i.e., VinVL and WESD, where VinVL’s bounding boxes are merged when they overlap. Since WESD makes bounding boxes that contain each connected segment of emotional stimuli, it tends to cover a large area. VinVL generates many small bounding boxes around objects, which coincide with the ground-truth bounding boxes.

Figure 11 compares WESD against CASNet. We find that CASNet tends to predict regions near the center of the image as emotional stimuli. This tendency is not surprising as the model is supervised by fixations from eye trackers and the image center seems to have a salient component. Meanwhile, WESD tends to find more relevant regions than

CASNet, at least for these examples (although because the difference in the scores between WESD and CASNet is small, the trend is not consistent for the APOLO test set).

In general, detecting emotional stimuli in artworks is still challenging as none of the three models perfectly spot the emotional stimuli in both Figures 10 and 11.

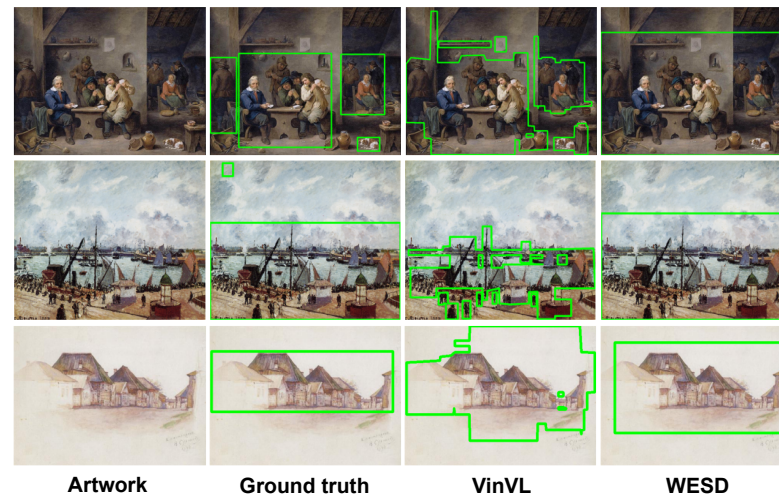


Figure 10. Examples of bounding box region detection. VinVL tends to distinguish objects exhaustively from the artwork, while WESD tends to find regions instead of certain objects.

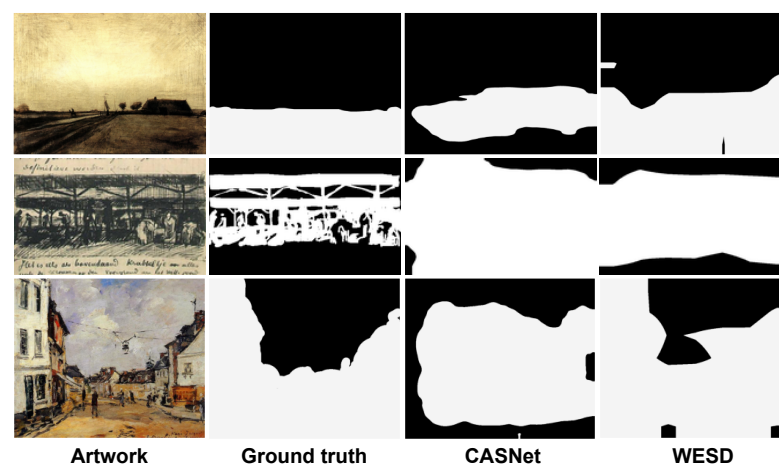


Figure 11. Examples of segmentation region detection. Compared to CASNet, WESD tends to find more related regions.

6.3. Emotion-Wise Analysis on Stimuli Detector

In this section, we analyze how well WESD predicts emotional stimuli maps for each emotion in one artwork. Specifically, we use the artworks in the test set of ArtEmis for this experiment, and we ask WESD to predict the emotional stimuli map for all of the emotions. We only evaluate WESD as it is the only baseline model that is able to predict emotional stimuli maps for each emotion and does not need utterances as a reference. Some results are shown in Figure 12. Through our experiments, we have the following observations:

- **Predictions focus on similar regions.** Although the WESD's predictions for each emotion are different, most of them focus on similar regions (e.g., the house and the pool in the first example and the people in the second example) in one artwork. The results could be reasonable as some regions in the artwork may play an essential role in evoking multiple emotions. We observe that such regions are also involved in the utterances.

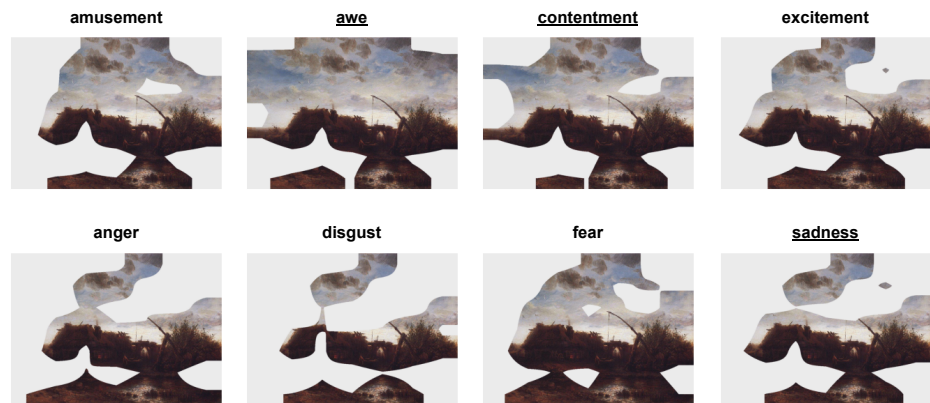
- **Awe and contentment tend to involve more regions.** Compared with other emotions, *awe* and *contentment* usually involve more regions, such as the whole sky in the first example, and the building and tree in the second example. These results may be related to the factor that the emotions of *awe* and *contentment* are usually evoked by wider scenery in the artwork.



awe: Nice depiction of scenery.

contentment: The bucolic setting is soothing and induces contentment. Pictures of nature relax me. The branch with the string hanging down looks like a tool that is used to make this place special.

sadness: It is dark and looks foreboding. It looks like it is either deserted or like someone lives there and does not have much money.



amusement: I like the little boy is playing with the dog that makes me happy.

awe: makes me feel like going back to simpler time. The choice in framing is beautifully done.

contentment: family and one member is playing with a dog. The children playing with a dog. The grandmother preparing food. The Father bringing in firewood. The cozy looking house.

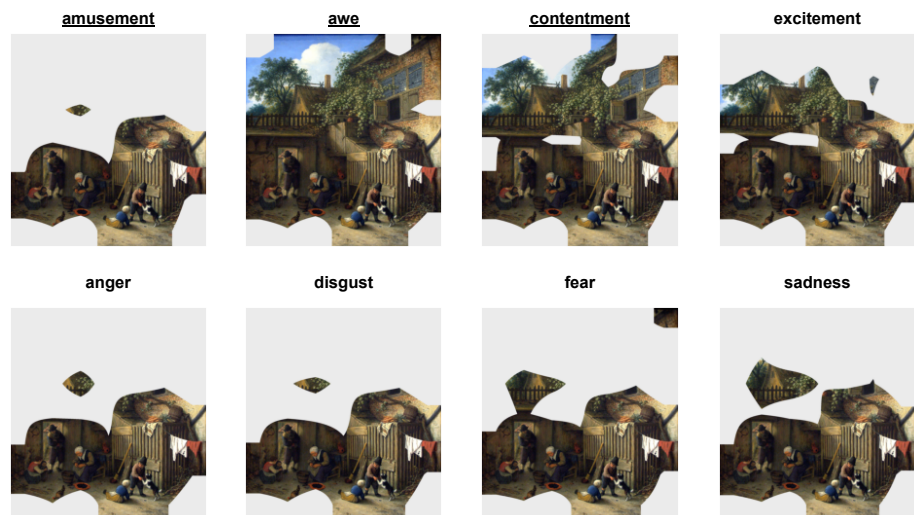


Figure 12. Examples of WESD’s prediction on eight different emotions. The texts on the left are the utterances from ArtEmis [11], which are not used during the prediction. The regions on the right are the predicted regions that evoke the corresponding emotions. The emotion tag on the right has an underline if this emotion appears on the left, i.e., has an annotation in ArtEmis.

7. Emotional Stimuli and Deep Generative Models

We consider emotional stimuli detection a task that may benefit training future deep generative models in generating emotional artworks, e.g., setting it as a loss function. In recent years, deep generative models, such as DALLÉ-2 [23] and Stable Diffusion [22], have demonstrated remarkable capabilities in producing high-quality images to users’ requirements. Such capacities also make these models popular in the artwork field, such as artwork generation [51] and editing [52].

In this section, we explore how much a popular deep generative model, Stable Diffusion [22], can handle emotions when generating artwork, and if our task and models can help improve its performance. To explore this, we randomly select 20 artists with one of their artworks in the APOLO dataset. Then, we make prompts by “The painting of [artwork name] by [artist name], produce [emotion]”. We use Stable Diffusion v1.5 [22] to generate

artworks for all combinations of 20 artworks and eight emotions, resulting in 160 generated artworks. Recently, DAAM [53] found that the aggregation of the cross-attention maps from Stable Diffusion can reveal the interpretation process of the model from prompts to images, i.e., reveal which parts of the image are related to a word in the prompt. We use DAAM to extract the internal attention map of *[emotion]*, which may indicate how Stable Diffusion interprets the emotion to the generated artwork.

We show some results in Figure 13. From the generated images, we find that Stable Diffusion can somehow generate artworks that can evoke certain emotions. However, from the internal attention map, we find that attention maps related to *[emotion]* are seldom focused. Instead, we observe that the attention sometimes focuses on the four corners of the artwork. These observations may indicate that it is still hard for Stable Diffusion to handle the relation between emotions and emotional stimuli. Compared to Stable Diffusion, WESD shows more concentration on the regions that are more related to the given emotions. The results may show a potential application of our work and WESD, to work as a guide and benefit Stable Diffusion in focusing on the emotional stimuli and generating more emotional artworks.

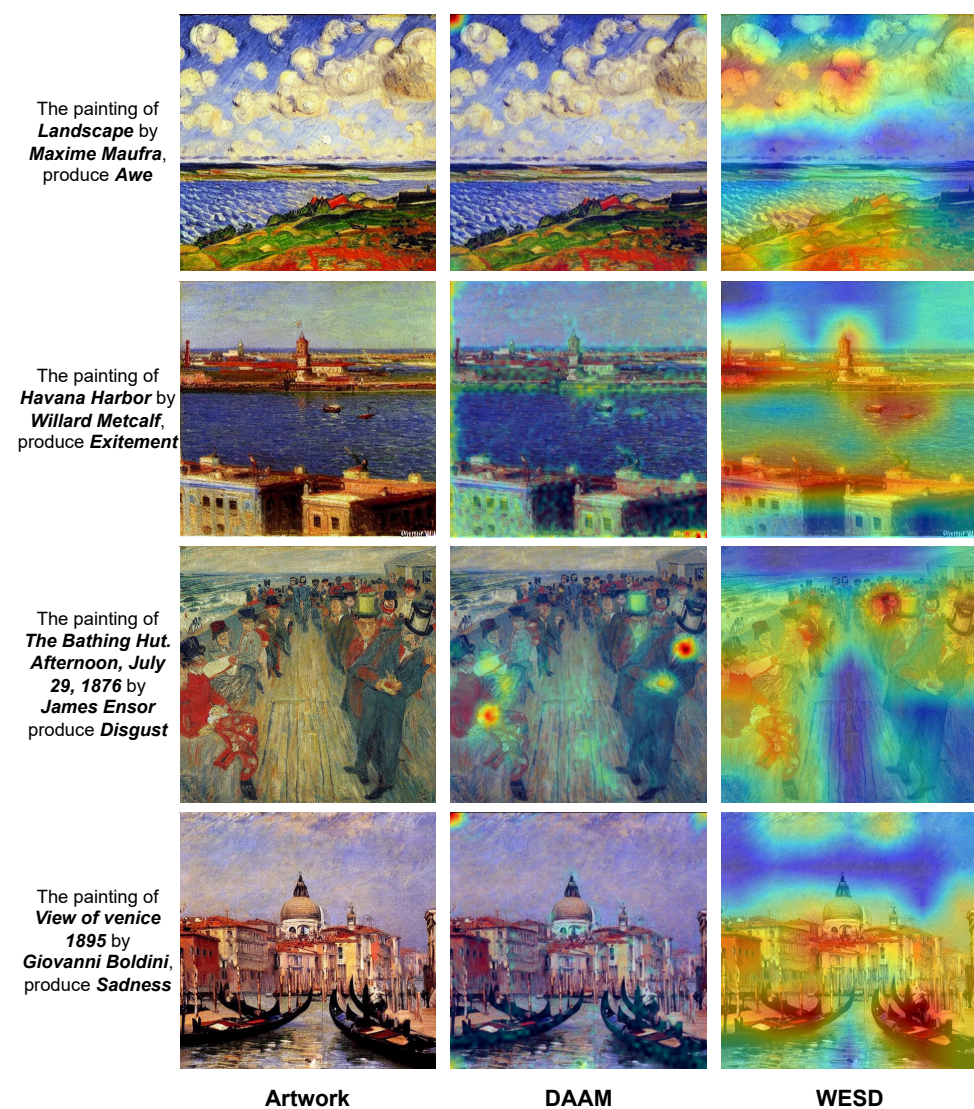


Figure 13. Examples of Stable Diffusion generated emotional artworks, the internal attention maps (DAAM), and WESD's predictions. The texts on the left are the prompts for Stable Diffusion. WESD could be useful for Stable Diffusion to generate emotional artwork by guiding the model to emphasize the emotional stimuli.

8. Limitations and Ethical Concerns

Our task is based on the appraisal theory of artworks and emotions [13,14]. Although this theory is reliable, it is continuously developing. We tried to remove inconsistent samples when constructing APOLO, as described in Section 4, but this may cause some domain gaps between our dataset and general artworks. Additionally, there are rising concerns about the ethical considerations of emotion recognition. As emotions are subjective and personal, trying to predict them with a machine learning model may be intrusive. We agree that emotion prediction could raise privacy issues and potential risks for model abuse. Being aware of this, we did our best to address these concerns proactively. In our experiments, we handled data responsibly and ensured that their use aligned with ethical standards. Additionally, we are planning to inform users of the inherent risks associated with our dataset and ensure they utilize it responsibly. Furthermore, we are prepared to take swift action, including freezing or deleting portions or the entirety of the dataset, if we identify any significant risks associated with its use. Through these measures, we hope to mitigate potential ethical risks and promote responsible usage of our research findings.

9. Conclusions

We introduced an emotional stimuli detection task that targets extracting regions from artworks that evoke emotions. For this task, we build a dedicated dataset, coined APOLO, with 6781 emotional stimuli in 4718 artworks for evaluation. We also provide APOLO with several baseline models to unveil the challenges in this task. Both qualitative and quantitative evaluations demonstrated that baseline models do not achieve a satisfactory performance, implying inherent difficulties in handling vague and abstract concepts of emotions. Furthermore, we explore how a deep generative model, Stable Diffusion, can handle emotions and emotional stimuli. We find that it is still hard for Stable Diffusion to understand and express emotions. We hope our work can bring inspiration to the fields of artwork analysis and visual emotion analysis.

Author Contributions: Conceptualization, T.C., N.G., L.L. and Y.N.; methodology, T.C., N.G., L.L. and Y.N.; software, T.C.; validation, T.C.; formal analysis, T.C.; investigation, T.C. and N.G.; resources, T.C. and N.G.; data curation, T.C. and N.G.; writing—original draft preparation, T.C.; writing—review and editing, T.C., N.G., L.L. and Y.N.; visualization, T.C.; supervision, N.G., L.L. and Y.N.; project administration, N.G. and Y.N.; funding acquisition, N.G. and Y.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JST CREST Grant No. JPMJCR20D3.

Data Availability Statement: Our data and methods are available at <https://github.com/Tianwei3989/apolo>.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

APOLO	<u>A</u> rtwork <u>P</u> rovoked em <u>O</u> tion Eva <u>L</u> uati <u>O</u> n
$BCE(\cdot)$	Binary cross-entropy
IoU	Intersection over union
$Pr@\theta$	Precision with IoU threshold θ
WESD	<u>W</u> eakly-supervised <u>E</u> motional <u>S</u> timuli <u>D</u> etection

Nomenclature

The following nomenclatures are used in this manuscript:

a	Artwork
B_w	Number of patches in the horizontal axes
B_h	Number of patches in the vertical axes

$\mathcal{D}_t$	Training set
$\mathcal{D}_v$	Validation set
$\mathcal{D}_e$	Test set
e	Emotion
$\mathcal{E}$	Set of emotions
$f_e(\cdot)$	Emotional stimuli detector
$F_g(\cdot)$	Fully-connected layer for embedding the whole artwork
$F_e(\cdot)$	Fully-connected layer to predict emotional stimuli score $\hat{y}_{ei}$
g_e	Binary emotion classifier for emotion e
i	Patch
K	Number of patches in artwork a
m_r	Map of region r
M_u	Emotional stimuli maps for utterance u
M_e	Emotional stimuli maps for emotion e
p	Point/location of segment s
P_w	Points related to phrase w
r	Region of artwork a
R	Regions in artwork a , a set of region r
s	Segment of artwork a
s_w	Segment of phrase w
$sim(\cdot, \cdot)$	Similarity score
$\hat{s}$	Segment predicted by model
θ	Threshold of IoU
u	Utterance
v_i	Feature of patch i
w	Phrase
W_u	Phrases in utterance u
y_{ei}	Emotional stimuli score of patch i with emotion e
$\hat{y}_{ei}$	Predicted emotional stimuli score of patch i with emotion e

References

1. Wilber, M.J.; Fang, C.; Jin, H.; Hertzmann, A.; Collomosse, J.P.; Belongie, S.J. BAM! The Behance Artistic Media Dataset for Recognition Beyond Photography. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017; pp. 1211–1220.
2. Bai, Z.; Nakashima, Y.; García, N. Explain Me the Painting: Multi-Topic Knowledgeable Art Description Generation. In Proceedings of the ICCV, Virtual Event, 11–17 October 2021; pp. 5402–5412.
3. Crowley, E.J.; Zisserman, A. The State of the Art: Object Retrieval in Paintings using Discriminative Regions. In Proceedings of the BMVC, Nottingham, UK, 1–5 September 2014.
4. Gonthier, N.; Gousseau, Y.; Ladjal, S.; Bonfait, O. Weakly supervised object detection in artworks. In Proceedings of the ECCV Workshops, Munich, Germany, 8–14 September 2018.
5. Mensink, T.; van Gemert, J.C. The Rijksmuseum Challenge: Museum-Centered Visual Recognition. In Proceedings of the ICMR, Dallas, TX, USA, 2 December 2014; p. 451.
6. Strezoski, G.; Worring, M. OmniArt: A Large-scale Artistic Benchmark. *ACM Trans. Multim. Comput. Commun. Appl.* **2018**, *14*, 88:1–88:21. [[CrossRef](#)]
7. Tonkes, V.; Sabatelli, M. How Well Do Vision Transformers (VTs) Transfer to the Non-natural Image Domain? An Empirical Study Involving Art Classification. In Proceedings of the ECCV Workshop, Tel Aviv, Israel, 23–27 October 2022; Volume 13801, pp. 234–250.
8. Reshetnikov, A.; Marinescu, M.V.; López, J.M. DEArt: Dataset of European Art. In Proceedings of the ECCV Workshop, Tel Aviv, Israel, 23–27 October 2022; Volume 13801, pp. 218–233.
9. Garcia, N.; Vogiatzis, G. How to Read Paintings: Semantic Art Understanding with Multi-modal Retrieval. In Proceedings of the ECCV Workshops, Munich, Germany, 8–14 September 2018; Volume 11130, pp. 676–691.
10. Garcia, N.; Ye, C.; Liu, Z.; Hu, Q.; Otani, M.; Chu, C.; Nakashima, Y.; Mitamura, T. A Dataset and Baselines for Visual Question Answering on Art. In Proceedings of the ECCV Workshops, Glasgow, UK, 23–28 August 2020; Volume 12536, pp. 92–108.
11. Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; Guibas, L.J. ArtEmis: Affective Language for Visual Art. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021; pp. 11564–11574.
12. Mohamed, Y.; Khan, F.F.; Haydarov, K.; Elhoseiny, M. It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2022; pp. 21231–21240.
13. Silvia, P.J. Emotional Responses to Art: From Collation and Arousal to Cognition and Emotion. *Rev. Gen. Psychol.* **2005**, *9*, 342–357. [[CrossRef](#)]

14. Cooper, J.M.; Silvia, P.J. Opposing Art: Rejection as an Action Tendency of Hostile Aesthetic Emotions. *Empir. Stud. Arts* **2009**, *27*, 109–126. [\[CrossRef\]](#)
15. Xenakis, I.; Arnellos, A.; Darzentas, J. The functional role of emotions in aesthetic judgment. *New Ideas Psychol.* **2012**, *30*, 212–226. [\[CrossRef\]](#)
16. Pelowski, M.; Akiba, F. A model of art perception, evaluation and emotion in transformative aesthetic experience. *New Ideas Psychol.* **2011**, *29*, 80–97. [\[CrossRef\]](#)
17. Yang, J.; Li, J.; Wang, X.; Ding, Y.; Gao, X. Stimuli-Aware Visual Emotion Analysis. *IEEE Trans. Image Process.* **2021**, *30*, 7432–7445. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Yang, J.; Gao, X.; Li, L.; Wang, X.; Ding, J. SOLVER: Scene-Object Interrelated Visual Emotion Reasoning Network. *IEEE Trans. Image Process.* **2021**, *30*, 8686–8701. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Yang, J.; She, D.; Lai, Y.K.; Rosin, P.L.; Yang, M.H. Weakly Supervised Coupled Networks for Visual Sentiment Analysis. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7584–7592.
20. Xu, L.; Wang, Z.; Wu, B.; Lui, S.S.Y. MDAN: Multi-level Dependent Attention Network for Visual Emotion Analysis. In Proceedings of the CVPR, Louisiana, LA, USA, 18–24 June 2022; pp. 9469–9478.
21. Yang, J.; Huang, Q.; Ding, T.; Lischinski, D.; Cohen-Or, D.; Huang, H. EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes. In Proceedings of the ICCV, Paris, France, 2–6 October 2023; pp. 20326–20337.
22. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In Proceedings of the CVPR, Louisiana, LA, USA, 18–24 June 2022; pp. 10674–10685.
23. Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; Chen, M. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* **2022**, arXiv:2204.06125.
24. Shen, X.; Efros, A.A.; Aubry, M. Discovering Visual Patterns in Art Collections with Spatially-Consistent Feature Learning. In Proceedings of the CVPR, Beach, CA, USA, 16–20 June 2019; pp. 9270–9279.
25. Lin, H.; Jia, J.; Guo, Q.; Xue, Y.; Huang, J.; Cai, L.; Feng, L. Psychological stress detection from cross-media microblog data using Deep Sparse Neural Network. In Proceedings of the ICME, Chengdu, China, 14–18 July 2014; pp. 1–6.
26. Wang, X.; Zhang, H.; Cao, L.; Feng, L. Leverage Social Media for Personalized Stress Detection. In Proceedings of the ACM MM, Seattle, WA, USA, 12–16 October 2020.
27. Truong, Q.T.; Lauw, H.W. Visual Sentiment Analysis for Review Images with Item-Oriented and User-Oriented CNN. In Proceedings of the ACM MM, Mountain View, CA, USA, 23–27 October 2017.
28. Truong, Q.T.; Lauw, H.W. VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. *Proc. Aaai Conf. Artif. Intell.* **2019**, *33*, 305–312. [\[CrossRef\]](#)
29. Zhao, S.; Yao, X.; Yang, J.; Jia, G.; Ding, G.; Chua, T.S.; Schuller, B.W.; Keutzer, K. Affective Image Content Analysis: Two Decades Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6729–6751. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Peng, K.C.; Sadovnik, A.; Gallagher, A.C.; Chen, T. Where do emotions come from? Predicting the Emotion Stimuli Map. In Proceedings of the ICIP, Phoenix, AZ, USA, 25–28 September 2016; pp. 614–618.
31. Fan, S.; Shen, Z.; Jiang, M.; Koenig, B.L.; Xu, J.; Kankanhalli, M.; Zhao, Q. Emotional Attention: A Study of Image Sentiment and Visual Attention. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7521–7531.
32. Liu, G.; Yan, Y.; Ricci, E.; Yang, Y.; Han, Y.; Winkler, S.; Sebe, N. Inferring Painting Style with Multi-Task Dictionary Learning. In Proceedings of the IJCAI, Buenos Aires, Argentina, 25–31 July 2015.
33. Ypsilantis, N.A.; Garcia, N.; Han, G.; Ibrahimi, S.; Van Noord, N.; Tolia, G. The met dataset: Instance-level recognition for artworks. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), San Diego, CA, USA, 8–14 December 2021.
34. Ruta, D.; Gilbert, A.; Aggarwal, P.; Marri, N.; Kale, A.; Briggs, J.; Speed, C.; Jin, H.; Faieta, B.; Filipkowski, A.; et al. StyleBabel: Artistic Style Tagging and Captioning. In Proceedings of the ECCV, Tel Aviv, Israel, 23–27 October 2022.
35. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the ACM MM, Firenze, Italy, 25–29 October 2010.
36. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the ICML, Virtual, 18–24 July 2021.
37. Kazemzadeh, S.; Ordonez, V.; Matten, M.A.; Berg, T.L. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In Proceedings of the EMNLP, Doha, Qatar, 25–29 October 2014.
38. Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.L.; Murphy, K.P. Generation and Comprehension of Unambiguous Object Descriptions. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 11–20.
39. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the NeurIPS, Las Vegas, NV, USA, 27–30 June 2019.
40. Lu, J.; Goswami, V.; Rohrbach, M.; Parikh, D.; Lee, S. 12-in-1: Multi-Task Vision and Language Representation Learning. In Proceedings of the CVPR, Seattle, WA, USA, 13–19 June 2020; pp. 10434–10443.
41. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proceedings of the ACL, Melbourne, Australia, 15–20 July 2018.
42. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. VinVL: Revisiting Visual Representations in Vision-Language Models. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021; pp. 5575–5584.

43. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
44. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
45. Fan, S.; Shen, Z.; Jiang, M.; Koenig, B.L.; Kankanhalli, M.S.; Zhao, Q. Emotional Attention: From Eye Tracking to Computational Modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1682–1699. [[CrossRef](#)] [[PubMed](#)]
46. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.P.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *40*, 834–848. [[CrossRef](#)] [[PubMed](#)]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the CVPR, Boston, MA, USA, 7–12 June 2015; pp. 770–778.
48. Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; Torr, P.H.S. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In Proceedings of the CVPR, Louisiana, LA, USA, 18–24 June 2022; pp. 18134–18144.
49. Xu, L.; Huang, M.H.; Shang, X.; Yuan, Z.; Sun, Y.; Liu, J. Meta Compositional Referring Expression Segmentation. In Proceedings of the CVPR, Tokyo, Japan, 15 June 2023; pp. 19478–19487.
50. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the ICLR, Toulon, France, 24–26 April 2017.
51. Wu, Y.; Nakashima, Y.; Garcia, N. Not Only Generative Art: Stable Diffusion for Content-Style Disentanglement in Art Analysis. In Proceedings of the ICMR, Thessaloniki, Greece, 12–15 June 2023; pp. 199–208.
52. Brooks, T.; Holynski, A.; Efros, A.A. InstructPix2Pix: Learning to Follow Image Editing Instructions. In Proceedings of the CVPR, Vancouver, Canada, 18–22 June 2023; pp. 18392–18402.
53. Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenettorp, P.; Lin, J.; Ture, F. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In Proceedings of the ACL, Toronto, ON, Canada, 9–14 July 2023; pp. 5644–5659.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.