# D-ViSA: A Dataset for Detecting Visual Sentiment from Art Images

Seoyun Kim, ChaeHee An, Junyeop Cha, Dongjae Kim, Eunil Park*
Sungkyunkwan University, Seoul, Korea
{seoyun.kim, anch0715, jycha95, bronze.ash}@g.skku.edu, eunilpark@skku.edu

## Abstract

*Detecting emotions evoked by art has been receiving great attention recently. Although previous works provide a variety of datasets consisting of art images and corresponding emotion labels, little attention has been paid to the continuous and dimensional characteristics of human emotions, especially in the domain of art. We propose a dataset for detecting visual sentiment from art images, D-ViSA, whose labels consist of both categorical and dimensional emotion labels which can be implemented in a wide range of visual sentiment analysis research regarding art. We compare several deep learning baselines in two specific tasks, single-feature, and multi-feature dimensional emotion regression. Our experiments lead to the conclusion that our dataset is plausible for both regression tasks with deep learning baselines. We assume that our dataset contributes to the field of artwork analysis and provides insights into human emotions evoked by art. The dataset is available at https://github.com/dxlabskku/D-ViSA*

## 1. Introduction

In recent years, visual sentiment recognition from images has been gaining more attention in the field of computer vision research [4, 19, 51]. This research area has a wide range of applications, ranging from general images including human face and body images [64, 71], and art images, which can include more abstract and aesthetic meanings [51, 57]. Among them, detecting emotions evoked by art is considered one of the most crucial and demanding research areas [1]. Art is a form of creative expression by humans, which aims to be admired, stimulate contemplation, and elicit emotional reactions [41]. Moreover, several prior studies have shown that the ability to evoke emotional responses through art is a desirable attribute, which can influence natural selection [3, 14, 17].

The primary intention behind the creation of most artworks is to deliberately provoke emotional reactions in viewers, who can often experience the emotions evoked by the artwork. Among various categories of artworks, abstract art stands out as a form originated by the sole purpose of eliciting emotional reactions through non-representational elements inherent in the artwork, such as, colors, lines, shapes, and textures [67]. Typically, abstract art lacks contextual details, which could potentially influence the emotions of observers [34], particularly challenging the examination of the emotion detection from it. Especially, abstract expressionism is known for its non-representational use of paint as a means of personal expression, emotions, inner experiences, and subconscious thoughts [21]. This movement aims to create diverse new visual languages and to pave the way for a broader concept of abstract art [5].

Thus, the intrinsic nature of art images can be regarded as distinct from that of natural images, which are often labeled based on their objective contents, focusing on the depicted objects or actions [8, 11]. Due to its diverse characteristics, the domain of art images can provide valuable data sources for defining semantically meaningful image analysis tasks. Moreover, it presents a challenging opportunity for neural networks to learn representations of a higher level of abstraction, considering the non-figurative characteristics [67]. To address the ambiguous and complex nature of human emotions, the dimensional emotion theory has gained prominence as a computational model for emotion recognition [2, 36]. It suggests that emotions can be represented by points in a continuous multidimensional space, rather than being solely categorized as sentiments (positive and negative) or emotion categories (e.g., excitement, fear). Within this framework, emotions can be described in terms of three dimensions: valence, arousal, and dominance [40, 44, 54, 55].

Valence refers to the emotional tone of an experience, where positive valence signifies a pleasant or happy feeling, while negative valence represents an unpleasant or sad feeling [53, 54, 56]. Arousal refers to the level of physiological and psychological activation associated with an emotional experience. The higher arousal emotions are typically intense and energizing, whereas the lower arousal emotions are calm and relaxing [53, 54, 56]. Dominance refers to the

*Corresponding author

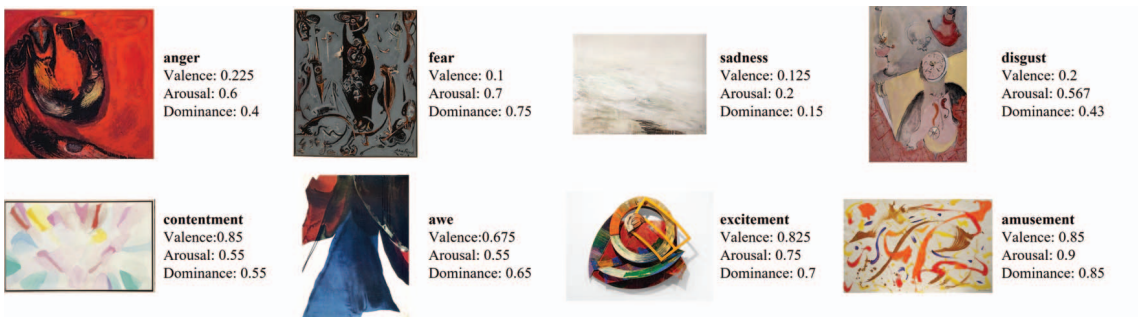


Figure 1. Examples with images and their corresponding categorical and VAD levels

degree of control or influence an emotional experience has over an individual, with highly dominant emotions getting a strong impact on behavior and thoughts, and less dominant emotions getting a weaker impact [37]. The aim of dimensional emotion detection is to provide a more nuanced understanding of the emotions conveyed by an image, moving beyond simple categorizations such as "happy" or "sad". Thus, recent researchers in computer vision tend to focus on continuous dimensional emotions, which can explain subtle, complex, and continuous affective behaviors.

We aim to enhance the understanding of the emotions evoked by art images. To achieve this, we present a dataset comprising abstract expressionism art images, which have been carefully annotated and validated with both categorical and dimensional emotion labels. Figure 1 shows a set of representative examples. Then, we attempt to verify whether the dimensional emotion detection tasks can be effectively performed using deep neural networks on this dataset. In addition to using simple images as input, we aim to enhance the efficiency of dimensional emotion detection by incorporating sparse categorical emotion label as an additional feature. Our study has the potential to provide notable insights into the emotional impact of artwork and contribute to the fields of artwork analysis and curation.

To summarize, our contributions are outlined as follows.

- We introduce *D-ViSA*, a dataset of visual sentiment evoked by art images annotated with dimensional emotion labels, as well as categorical emotion labels.
- To the best of our knowledge, it is one of the first efforts providing all three-dimensional emotion labels coupled with art images, validating its plausibility for dimensional emotion prediction.
- Leveraging *D-ViSA*, we present a novel deep learning framework, which incorporates both images and their categorical labels as input features, enhancing the performance on the dimensional emotion prediction task.

## 2. Related Work

Visual sentiment analysis and detection have garnered attention due to their relevance in not only computer vision but psychology studies. By incorporating such datasets and dimensional labels, studies have been conducted to deal with more detailed and abundant human emotions.

### 2.1. Employing Dimensional Emotion

To gain a more comprehensive representation on the affective states of the subjects in human facial expression research [7, 25, 52] and general-topic image research [26, 35, 42, 49, 73], researchers have put efforts to optimize the annotation process based on the characteristics of given data. One notable example is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [7], comprised total of 10,000 videos featuring ten speakers, annotated by a minimum of three annotators. In addition to the emotion labels following Ekman's definition, IEMOCAP offers rich primitive attributes: valence, arousal, and dominance. REmote COLlaborative and Affective (RECOLA) database [52] utilizes a comprehensive dataset that includes audio, video, electrocardiogram (ECG), and electrodermal activity (EDA) modalities. The dataset consists of 9.5 hours of recordings from 46 French-speaking participants engaged in a collaborative task during a video conference. For each recording, the V-A annotations are annotated by six French-speaking annotators.

Apart from datasets focusing on facial expression, the Nencki Affective Picture System (NAPS) [35] is a collection of 1,356 high-quality photographs consists of general objects and landscapes. A total of 204 participants, primarily from Europe, rated these images using a 9-point bipolar semantic sliding scale to measure VA and approach-avoidance dimensions. EMOTIC [26] is another image database with people in real-life environments. It combines images from three online sites [33, 75] and Google, and employs annotators from crowdsourcing. The dataset includes 18,316 images and annotations covering 26 emotion categories and the 10-point VAD dimensions.

By incorporating such datasets and dimensional labels, studies have been conducted to deal with more detailed and abundant human emotions. Some studies using general images including social media images are conducted [42, 73], improving not only the understanding on emotion analysis

but also network architectures used for studies. Moreover, various attempts are made for continuous human emotion recognition tasks [13, 29, 59, 63, 64].

### 2.2. Art Emotion Dataset

The paintings can elicit emotional responses in viewers, enabling researchers to leverage the presentation of paintings to induce specific emotional states, and to investigate on the associated cognitive and physiological processes. Therefore, painting datasets for emotional research assumes critical significance, with WikiArt [41], ArtEmis [1], abstract paintings [34] emerging as widely used resources in this domain of research [6, 9, 51].

Both WikiArt emotions dataset [41] and ArtEmis dataset [1] consist of a collection of more than 4,000 art images sourced from the WikiArt database. Also, both are annotated through crowdsourcing. The former is labeled with 19 emotions derived from previous studies [18, 39, 43, 47, 50, 60, 61]. The images of ArtEmis are also labeled based on previous studies [34, 68, 69, 72], annotators identified the dominant emotion they perceived from the given artwork and provided a grounded explanation for the choice. For abstract paintings dataset [34] which is proposed to explore and advance techniques for extracting and integrating low-level features that represent the emotional content of images. A collection of 228 abstract paintings are peer rated through web-survey by 230 annotators, categorized based on Mikel's emotions [38].

Hence, partial attempts to adopt the utilization of continuous labels have also been done in the realm of affective computing studies for art. Notably, two prominent positive-negative emotional valence datasets that have been extensively utilized in research are the Museum of Modern and Contemporary Art of Trento and Rovereto dataset (MART) [57, 67], and the deviantArt dataset (devArt) [57].

The MART comprises 500 abstract paintings carefully selected from the art collection of the MART museum in Rovereto. In contrast, the devArt includes 500 abstract paintings sourced from amateur artists in the online art community. To measure the positive-negative scale values in the MART, a total of 100 annotators provided the scoring for the positive-negative scale using both Absolute Scale scores and Relative scores. Conversely, the devArt solely utilized Relative scores for positive-negative measurement.

Multiple datasets have been employed in visual sentiment studies, accompanied by the development of numerous sentiment analysis models. Despite the availability of emotion labels for human facial expressions or social media images, the scarcity of datasets containing both categorical emotion labels and VAD levels specifically for art painting images is evident. Furthermore, along with the attempts to employ both categorical and dimensional emotion labels to form better affective recognition using textual data [15, 46], deploying both resources while modeling with image data could be taken into account.

## 3. D-ViSA

*D-ViSA*, a collection of 2,782 abstract art images accompanied by two types of labels indicating categorical emotions and dimensional emotions, is proposed. The procedures for building this dataset are presented as follows.

### 3.1. Data Collection and Annotation

We built a dataset of abstract art images, distinguishing those from social media or facial images. To achieve this, we collected abstract expressionist art images from the WikiArt [41]. This database is a collection of publicly available art images, which is organized by over 250,000 images originating from different art movements and genres. For our dataset, we carefully selected a subset of 2,782 abstract expressionist art images from the WikiArt database.

To ensure the dataset was labeled with a high level of credibility, we formed a team of three annotators, who possessed relevant credentials [28]. All annotators had backgrounds in Fine Arts, and completed both undergraduate and master's degrees in this field. While emotions can be influenced by a number of factors, they are considered to be high-level cognitive processes, which exhibit a certain level of stability and universality across individuals and cultures [45]. It allowed us to generalize the emotional labels based on the opinions of these selected individuals.

For the categorical emotion labels, the annotators were provided with a set of instructions [31]. They were given a set of eight basic emotion categories (excitement, amusement, contentment, awe, sadness, anger, disgust, and fear) suggested by Mikel [38]. They were then asked to select one or more emotional labels based on the feelings evoked by viewing each given image. Subsequently, to assign dimensional emotion labels, the experimenter checked that annotators were equipped with comprehensive instructional materials considering dimensional emotions, ensuring their understanding of the concept. They were provided with the NRC-VAD Lexicon [40], which includes a range of emotions and their corresponding Valence-Arousal-Dominance (VAD) labels, as well as Mikel's emotion wheel as supplementary material. Using them as notable references, the annotators were instructed to rate the VAD levels for each art image on a scale ranging from 0 to 1 [40].

The annotation procedures were carried out via online websites. They presented each art image along with textual information including the author, title, and published year. It allowed the annotators to enter one or more categorical emotion labels and to rate the VAD levels for each image through designated input boxes. The questionnaire items associated with each input box were presented as follows:

1. **Categorical emotion**: Please select one or more emotions evoked by the displayed artwork from the following options: excitement, amusement, contentment, awe, sadness, anger, disgust, and fear.

2. **Valence**: On a scale from 0 to 1, please rate the level of positivity or negativity of the emotion evoked by the artwork (0: negative emotion, 1: positive emotion).

3. **Arousal**: On a scale from 0 to 1, please rate the level of physical excitement evoked by the artwork (0: no excitement, 1: the greatest excitement).

4. **Dominance**: On a scale from 0 to 1, please rate the level of perceived control or loss of control evoked by the artwork (The higher level suggests a stronger sense of dominance or loss of control).

### 3.2. Data Validation

A total of 2,782 art images were sequentially displayed and labeled by three annotators. As a result, we simultaneously obtained three categorical emotion labels and VAD levels on each image. To finalize the labels/levels, we aggregated the ratings from all annotators as follows:

- **Categorical emotion label**: The final categorical emotion label for each art image was determined by selecting the most frequently occurring labels.

- **VAD levels**: The final values of the Valence, Arousal, and Dominance dimensions were determined by computing the mean value of the respective ratings provided by the annotators.

In the following three cases, additional inspection was conducted. The first case involved instances where the categorical emotion labels could not be determined by utilizing the most frequently occurring labels, due to conflicting annotations (e.g., one contentment, one excitement, and one amusement label from the three annotators). The second case occurred when the aggregated combination of categorical emotion labels exhibited a mixture of highly positive emotions (e.g., excitement) and strongly negative emotions (e.g., anger). The last case is some instances, where errors were identified in the procedures, such as missing labels or inappropriate ranges in the VAD levels. In such cases, further inspection was carried out to examine the final single categorical emotion label.

To validate the categorical labels, we followed the similar approach done on several prior works on image classification tasks [58], by recruiting three inspectors, who had extensive experience in data annotations and labeling procedures. First, each inspector was asked to pick a single categorical label among the aggregation of all categorical emotion inputs provided by the annotators for each image. The selected single categorical emotion became the final categorical emotion label ($e_{final}$) for the corresponding image. For dimensional labels, Figure 2 shows the procedures of the finalized VAD levels ($VAD_{final}$).

### 3.3. Descriptive Statistics of Datasets

Table 1 presents the descriptive statistics of the dataset, providing an overview of the distribution of categorical emotion labels with VAD levels. *fear* (584, 21%) label was the most frequently employed label, followed by *contentment* label (441, 16%), while *anger* (105, 4%) label was the least annotated label among the eight emotion categories.

In Figure 3, we present the correlation among the VAD dimensions, which are presented by calculating Pearson's correlation coefficient ($r$) based on the VAD levels assigned to the art images [12]. The correlation matrix reveals a strong positive correlation between arousal and dominance ($r$ = 0.84). However, in contrast, the correlations between valence and arousal, as well as valence and dominance, are low ($r$ = -0.06 and 0.04, respectively).

Figure 4 shows the correlation results among the average VAD levels for the emotions. The correlation values between negative (*anger, sadness, fear*) and positive (*contentment, amusement, and excitement*) emotions tend to be highly negative, approaching -1, as these emotions represent opposite ends of the emotional spectrum. However,

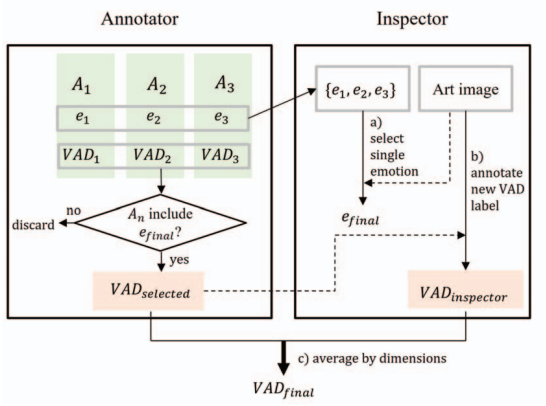


Figure 2. Summary of the annotation procedures (VAD levels)

| | # of samples (%) | Mean (Standard deviation) V | A | D |
|---|---|---|---|---|
| *fea* | 587 (21%) | .20 (.10) | .60 (.18) | .53 (.19) |
| *con* | 441 (16%) | .68 (.11) | .41 (.20) | .40 (.19) |
| *awe* | 376 (14%) | .60 (.17) | .63 (.19) | .57 (.21) |
| *sad* | 375 (13%) | .24 (.10) | .49 (.19) | .44 (.21) |
| *exc* | 344 (12%) | .65 (.13) | .56 (.14) | .53 (.16) |
| *amu* | 305 (11%) | .67 (.11) | .53 (.15) | .48 (.17) |
| *dis* | 249 (9%) | .23 (.10) | .55 (.17) | .50 (.19) |
| *ang* | 105 (4%) | .27 (.11) | .58 (.17) | .51 (.15) |
| Total | 2,782 (100%) | .45 (.24) | .54 (.19) | .50 (.20) |

Table 1. Overall statistics of each emotion category

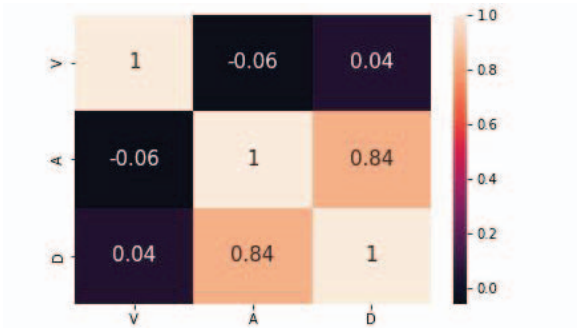


Figure 3. Pearson's correlation coefficients among VAD levels

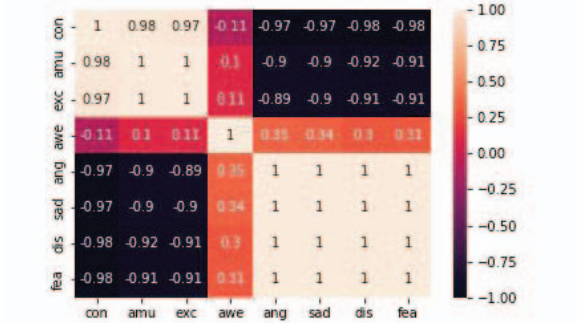


Figure 4. Pearson's correlation coefficients of mean VAD levels between eight emotion categories

the correlation between *awe* and other emotions varies between -0.11 and 0.35. It is because *awe* is one of the emotions, which can encompass both negative and positive feelings [41]. We also present several examples with their corresponding VAD levels (Figure 1).

## 4. Detecting Emotions Evoked by Art

Our approach focuses on validating the plausibility of our dataset, *D-ViSA*, by employing deep learning models to assess their ability to accurately capture semantic features and detect emotions from art images. For achieving this purpose, we train a three-dimensional image regression model so that it could predict true VAD levels of our dataset from the corresponding artwork. To predict VAD levels from art images, we implemented two frameworks: a single-feature regression framework for utilizing only the image as an input, and a multi-feature regression framework for incorporating both the image and its emotion label. In both frameworks, Mean Squared Error (MSE), which is a commonly used loss function by computing the discrepancy between predicted values and ground truth labels, was implemented as the loss function for regression. It enables to update the gradients during the training procedures. We provide more detailed information on the experimental setup and implementation of two frameworks as follows:

$$L_{\text{MSE}}(y_{dim}, \hat{y}_{dim}) = \frac{1}{n}\sum_{i=1}^{n}(y^i_{dim} - \hat{y}^i_{dim})^2 \qquad (1)$$

where $n$ refers to 3 (e.g., valence, arousal, and dominance). In addition, $y_{dim}$ and $\hat{y}_{dim}$ represent true and predicted VAD levels, respectively.

### 4.1. Single-feature Dimensional Emotion Prediction

In the single feature dimensional emotion regression task, the final layer comprises a three-dimensional fully-connected layer. Each dimension of this layer represents valence, arousal, and dominance ranging from 0 to 1. The model extracts $n$-dimensional image features from each input image. These image features capture the relevant information necessary for predicting the dimensional emotions. The extracted features are then passed to the final fully-connected layer, which produces a three-dimensional vector representing the predicted VAD levels.

### 4.2. Multi-feature Dimensional Emotion Prediction

In the multi-feature dimensional emotion regression task, the final layer of the model also consists of a three-dimensional fully-connected layer, with each dimension representing valence, arousal, and dominance. In addition to extracting the $n$-dimensional image feature using each baseline model, the multi-feature model incorporates an additional input feature: the one-hot encoded categorical emotion label corresponding to the input image. This categorical emotion label feature is extended from 8 dimensions to $n$ dimensions using a single fully-connected layer, aligning its size with the extracted image feature. The $n$-dimensional image and the extended categorical emotion label features are then concatenated to create a $2n$-dimensional final feature representation. The concatenated feature is passed into the final fully-connected layer, which outputs a three-dimensional vector representing the predicted VAD levels.

## 5. Experiments

The experiments were conducted with a single RTX A6000 48GB GPU and Python 3.8. We used Pytorch [48] to implement our models. The dataset was split into train and test sets (8:2). Moreover, 10% of the train set was further split and employed as a validation set for monitoring the training performance of a model. Thus, the samples of training, validation, and test sets are counted as 2,002 (72%), 223 (8%), and 557 (20%), respectively. To enhance the training procedures, data augmentation techniques, horizontal flip, and random-resized crop [10], were applied for the training set. We implemented the Adam optimizer [24] with a learning rate set to 0.01. We employed a batch size of 32 and trained the models for 50 epochs [30, 22]. Throughout the training session, we periodically saved the model and finally chose the best-performing one with the lowest validation loss. To evaluate the performance of the VAD regression, we calculated a Pearson's correlation coefficient

| | Single-feature regression | | | | | | Multi-feature regression | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test | | | Validation | | | Test | | | Validation | | |
| | V | A | D | V | A | D | V | A | D | V | A | D |
| VGG16 [62] | - | 0.19 | 0.23 | 0.14 | 0.17 | 0.22 | 0.87 | 0.36 | 0.31 | 0.89 | 0.43 | 0.30 |
| AlexNet [27] | - | - | - | - | - | - | 0.87 | 0.34 | 0.30 | 0.88 | 0.43 | 0.31 |
| ResNet-50 [20] | 0.50 | 0.24 | 0.24 | 0.45 | 0.21 | 0.25 | 0.87 | 0.34 | 0.34 | 0.88 | 0.46 | 0.39 |
| PDANet [74] | 0.53 | 0.27 | 0.27 | 0.56 | 0.23 | 0.31 | 0.89 | 0.41 | 0.36 | 0.89 | 0.40 | 0.38 |
| ViT [16] | 0.39 | 0.08 | 0.09 | 0.33 | - | 0.17 | 0.48 | 0.26 | 0.18 | 0.49 | 0.20 | 0.26 |
| Zhang et al. [70] | 0.35 | 0.21 | 0.18 | 0.30 | 0.19 | 0.25 | 0.85 | 0.34 | 0.22 | 0.86 | 0.33 | 0.25 |

Table 2. The overall performance of single-feature regression and multi-feature regression models on *D-ViSA*. Pearson's correlation coefficients were derived to evaluate the performance.

($r$), which is known for its strength in providing stable evaluation even with different scaling factors and various training objectives, between the predicted and actual VAD levels, following the methodology outlined by several prior research [46, 65, 66]. The formula is presented as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2)$$

where $x \in X$, $y \in Y$. In detail, $X$ refers to "*predicted VAD levels*", while $Y$ "*refers to ground truth VAD levels*" [32]. $n$, $\bar{x}$, and $\bar{y}$ indicate the number of data points, sample mean of $X$, and sample mean of $Y$, respectively.

We evaluate our dataset using several deep-learning models, as common baselines: **VGG16** [62], **AlexNet** [27], and **ResNet-50** [20], which were fine-tuned with the initial pre-trained weights and without parameter freezing, by using 448×448 resized images [23]. Moreover, we employed three state-of-the-art models as follows: PDANet [74], **Vision Transformer (ViT)** [16], and **Zhang et al.** [70]. These models were trained with the objective of accurately predicting the three-dimensional VAD labels.

## 6. Results

Table 2 shows the Pearson's correlation coefficients for each dimension of predicted VAD levels by the six deep learning baseline models for two experimental tasks (single-feature and multi-feature regression tasks). The overall performance indicates that *D-ViSA* is sufficiently implementable for dimensional emotion detection tasks, achieving considerable evaluation scores. In the single-feature regression task, **PDANet**, one of the state-of-the-art models, demonstrated the highest performance on both the test and validation sets among the baselines.

In the multi-feature regression task, **PDANet** exhibited the highest performance in predicting Valence, Arousal, and Dominance in the test set, and Valence in the validation set. For Arousal and Dominance in the validation set, **ResNet-50** achieved the best scores of 0.46 and 0.39, respectively. Generally, the multi-feature regression models outperformed the single-feature regression models, indicating that the proposed multi-feature regression framework can generate better representations capturing affective features from both art images and categorical emotion labels.

## 7. Conclusion

We present *D-ViSA*, a dataset for visual sentiment derived from art images. *D-ViSA* comprises a total of 2,782 abstract expressionism art images, accompanied by their corresponding single categorical and three-dimensional emotion labels. The categorical emotion labels are based on Mikel's eight emotion categories. Then, each art image in the dataset is finally assigned to one of these categorical emotion labels with three-dimensional emotion levels.

We conducted two VAD regression tasks (single-feature and multi-feature) with several baselines using art images input to predict three-dimensional VAD levels. The results obtained from these experiments demonstrated reasonable performance, indicating that *D-ViSA* is well-suited for dimensional emotion detection tasks. Furthermore, we observed that incorporating an additional categorical emotion feature into the regression task improved overall performance across the different deep-learning baselines.

These findings suggest that the combination of categorical and dimensional emotion features enhances the ability to accurately recognize and predict emotional characteristics in art images. Thus, *D-ViSA* offers a valuable resource for conducting dimensional emotion detection research in the field of art analysis.

We believe our dataset is a valuable resource for researchers who are interested in exploring the relationship between art and emotions using both categorical and dimensional emotion labels. Its availability may broaden the visual sentiment research domains, leading to new insights into how artwork evokes emotions and how these emotions are related to different dimensions of affective experiences.

## Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) program(IITP-2020-0-01816) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation).

## References


[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In Proc. of CVPR '21, pages 11569–11579, 2021. 1, 3

[2] Hyeongjin Ahn and Eunil Park. Motivations for user satisfaction of mobile fitness applications: An analysis of user experience based on online review comments. Humanities and Social Sciences Communications, 10(1):1–7, 2023. 1

[3] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In Proc. of NAACL-HLT '19, pages 370–379, 2019. 1

[4] Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe. Recognizing emotions from abstract paintings using non-linear matrix completion. In Proc. of CVPR '16, pages 5240–5248, 2016. 1

[5] David Anfam and Tate Gallery Liverpool. American Abstract Expressionism: Experiencing and Envisioning the City, volume 1. Liverpool University Press, 1993. 1

[6] Digbalay Bose, Krishna Somandepalli, Souvik Kundu, Rimita Lahiri, Jonathan Gratch, and Shrikanth Narayanan. Understanding of emotion perception from art. arXiv preprint arXiv:2110.06486, 2021. 3

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42:335–359, 2008. 2

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In Proc. of CVPR '15, pages 961–970, 2015. 1

[9] Ming Chen, Lu Zhang, and Jan P Allebach. Learning deep features for image emotion classification. In Proc. of ICIP '15, pages 4491–4495. IEEE, 2015. 3

[10] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. https://arxiv.org/abs/2001.04086, 2020. 5

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015. 1

[12] Yanyan Chi and Eunil Park. Counterattacking long videos: exploring the characteristics of popular instant videos and roles of producers and viewers. Library Hi Tech, 41(3):694–710, 2023. 4

[13] Dong Yoon Choi and Byung Cheol Song. Semi-supervised learning for facial expression-based emotion recognition in the continuous domain. Multimedia Tools and Applications, 79(37-38):28169–28187, 2020. 3

[14] Stephen Davies. The artful species: Aesthetics, art, and evolution. OUP Oxford, 2012. 1

[15] Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. Mixing and matching emotion frameworks: Investigating cross-framework transfer learning for dutch emotion detection. Electronics, 10(21):2643, 2021. 3

[16] 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 Proc. of ICLR '20, pages 1–21, 2020. 6

[17] Denis Dutton. The art instinct: Beauty, pleasure, & human evolution. Oxford University Press, USA, 2009. 1

[18] Paul Ekman. An argument for basic emotions. Cognition & emotion, 6(3-4):169–200, 1992. 3

[19] Maurice Gerczuk, Shahin Amiriparian, Sandra Ottl, and Bjorn W Schuller. Emonet: A transfer learning framework for multi-corpus speech emotion recognition. IEEE Transactions on Affective Computing, 14(02):1472–1487, 2023. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proc. of CVPR '16, pages 770–778, 2016. 6

[21] Barbara Hess. Abstract expressionism. Taschen, 2016. 1

[22] Syjung Hwang, Jina Kim, Eunil Park, and Sang Jib Kwon. Who will be your next customer: A machine learning approach to customer return visits in airline services. Journal of Business Research, 121:121–126, 2020. 5

[23] Honggeun Ji, ChaeHee An, Minyoung Lee, Jufeng Yang, and Eunil Park. Fused deep neural networks for sustainable and computational management of heat-transfer pipeline diagnosis. Developments in the Built Environment, 14:100144, 2023. 6

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 5

[25] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. Image and Vision Computing, 65:23–36, 2017. 2

[26] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In Proc. of CVPR '17, pages 1667–1675, 2017. 2

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012. 6

[28] Jinmo Lee and Eunil Park. D-hrsp: Dataset of helpful reviews for service providers. Telematics and Informatics, 82:102001, 2023. 3

[29] Kunyoung Lee, Seunghyun Kim, and Eui Chul Lee. Fast and accurate facial expression image classification and regression method based on knowledge distillation. Applied Sciences, 13(11):6409, 2023. 3

[30] Minyoung Lee, Honggeun Ji, and Eunil Park. Deepaup: A deep neural network framework for abnormal underground heat transport pipelines. IEEE Transactions on Automation Science and Engineering, 2023. 5

[31] Seungpeel Lee, Jina Kim, Dongjae Kim, Ki Joon Kim, and Eunil Park. Computational approaches to developing the

implicit media bias dataset: Assessing political orientations of nonpolitical news articles. Applied Mathematics and Computation, 458:128219, 2023. 3

[32] SangEun Lee, Chaeeun Ryu, and Eunil Park. Osanet: Object semantic attention network for visual sentiment analysis. IEEE Transactions on Multimedia, 2022. 6

[33] 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 Proc. of ECCV '14, pages 740–755. Springer, 2014. 2

[34] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In Proc. of ACM MM '10, pages 83–92, 2010. 1, 3

[35] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. Behavior research methods, 46:596–610, 2014. 2

[36] Stacy Marsella and Jonathan Gratch. Computationally modeling human emotion. Communications of the ACM, 57(12):56–67, 2014. 1

[37] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology, 14:261–292, 1996. 2

[38] Mikels, J. A. et al. Emotional category data on images from the international affective picture system. Behavior research methods, 37(4):626–630, 2005. 3

[39] Keith Millis. Making meaning brings pleasure: the influence of titles on aesthetic experiences. Emotion, 1(3):320, 2001. 3

[40] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In Proc. of ACL '18, pages 174–184, 2018. 1, 3

[41] Saif Mohammad and Svetlana Kiritchenko. Wikiart emotions: An annotated dataset of emotions evoked by art. In Proc. of LREC '18, pages 1225–1238, 2018. 1, 3, 5

[42] Sharoon Nasim, Mahnoor Rehan, and Nosheen Sabahat. Emotional understanding of an image by applying high-level concepts on image parts. In Proc. INMIC '20, pages 1–5. IEEE, 2020. 2

[43] Pinchas Noy and Dorit Noy-Sharav. Art and emotions. International journal of applied psychoanalytic studies, 10(2):100–107, 2013. 3

[44] Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. The measurement of meaning. University of Illinois press, 1957. 1

[45] Li-Chen Ou, M Ronnier Luo, Andrée Woodcock, and Angela Wright. A study of colour emotion and colour preference. part i: Colour emotions for single colours. Color Research & Application, 29(3):232–240, 2004. 3

[46] Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. Dimensional emotion detection from categorical emotion. In Proc. of EMNLP '21, pages 4367–4380, 2021. 3, 6

[47] W Gerrod Parrott. Emotions in social psychology: Essential readings. psychology press, 2001. 3

[48] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. http://note.wcoder.com/files/ml/automatic_differentiation_in_pytorch.pdf, 2017. 5

[49] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In Proc. of CVPR '15, pages 860–868, 2015. 2

[50] Robert Plutchik. A general psychoevolutionary theory of emotion. In Theories of emotion, pages 3–33. Elsevier, 1980. 3

[51] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multilevel deep representations for image emotion classification. Neural processing letters, 51:2043–2061, 2020. 1, 3

[52] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In Proc. of FG '13, pages 1–8. IEEE, 2013. 2

[53] James A Russell. Affective space is bipolar. Journal of personality and social psychology, 37(3):345–356, 1979. 1

[54] James A Russell. A circumplex model of affect. Journal of personality and social psychology, 39(6):1161–1178, 1980. 1

[55] James A Russell. Core affect and the psychological construction of emotion. Psychological review, 110(1):145, 2003. 1

[56] James A Russell, Maria Lewicka, and Toomas Niit. A cross-cultural study of a circumplex model of affect. Journal of personality and social psychology, 57(5):848, 1989. 1

[57] Andreza Sartori, Victoria Yanulevskaya, Almila Akdag Salah, Jasper Uijlings, Elia Bruni, and Nicu Sebe. Affective analysis of professional and amateur abstract paintings using statistical analysis and art theory. ACM Transactions on Interactive Intelligent Systems, 5(2):1–27, 2015. 1, 3

[58] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 8430–8439, 2019. 4

[59] Jinrui Shen, Jiahao Zheng, and Xiaoping Wang. Mmtransmt: A framework for multimodal emotion recognition using multitask learning. In Proc. of ICACI '21, pages 52–59. IEEE, 2021. 3

[60] Paul J Silvia. Emotional responses to art: From collation and arousal to cognition and emotion. Review of general psychology, 9(4):342–357, 2005. 3

[61] Paul J Silvia. Looking past pleasure: anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. Psychology of Aesthetics, Creativity, and the Arts, 3(1):48, 2009. 3

[62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014. 6

[63] Minh Tran and Mohammad Soleymani. A pre-trained audio-visual transformer for emotion recognition. In Proc. of ICASSP '22, pages 4698–4702. IEEE, 2022. 3

[64] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing, 11(8):1301–1309, 2017. 1, 3

[65] Kexin Wang, Zheng Lian, Licai Sun, Bin Liu, Jianhua Tao, and Yin Fan. Emotional reaction analysis based on multi-label graph convolutional networks and dynamic facial expression recognition transformer. In Proc. of MuSe '22, pages 75–80, 2022. 6

[66] Shangfei Wang, Longfei Hao, and Qiang Ji. Knowledge-augmented multimodal deep regression bayesian networks for emotion video tagging. IEEE Transactions on Multimedia, 22(4):1084–1097, 2019. 6

[67] Victoria Yanulevskaya, Jasper Uijlings, Elia Bruni, Andreza Sartori, Elisa Zamboni, Francesca Bacci, David Melcher, and Nicu Sebe. In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. In Proc. of ACM MM '12, pages 349–358, 2012. 1, 3

[68] Victoria Yanulevskaya, Jan C van Gemert, Katharina Roth, Ann-Katrin Herbold, Nicu Sebe, and Jan-Mark Geusebroek. Emotional valence categorization using holistic image features. In Proc. of ICIP '08, pages 101–104. IEEE, 2008. 3

[69] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In Proc. of AAAI '16, pages 308–314, 2016. 3

[70] Jing Zhang, Xinyu Liu, Mei Chen, Qi Ye, and Zhe Wang. Image sentiment classification via multi-level sentiment region correlation analysis. Neurocomputing, 469:221–233, 2022. 6

[71] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Multimodal deep convolutional neural network for audio-visual emotion recognition. In Proc. of SIGIR '16, pages 281–284, 2016. 1

[72] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In Proc. of ACM MM '14, pages 47–56, 2014. 3

[73] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression. In Proc. of ACM MM '19, pages 192–201, 2019. 2

[74] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In Proc. of AAAI '20, volume 34, pages 303–311, 2020. 6

[75] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision, 127:302–321, 2019. 2