Emotion Aggregation in Artistic Image Analysis: Effects of Label Distribution Learning

Ryuichi Takahashi[1], **Yuta Sasaki**[2], **Yuhki Shiraishi**[3], **Jianwei Zhang**[1]
[1]Iwate University {g0323115, zhang}@iwate-u.ac.jp
[2]Institute of Science Tokyo yubo1336@lr.pi.titech.ac.jp
[3]Tsukuba University of Technology yuhkis@a.tsukuba-tech.ac.jp

Abstract

This paper addresses the challenges of modeling human emotional responses to artwork through an exploration of Label Distribution Learning (LDL). We introduce Progressive Label Distribution Transition (PLDT), a novel framework that bridges the gap between traditional One-hot encoding and LDL by implementing gradual transitions between these paradigms. To evaluate our approach, we propose TESA (Thresholded Emotion Set Accuracy), a comprehensive evaluation framework. Our threshold-based analysis reveals new insights into how these methods balance prediction confidence and emotional multiplicity in artwork perception. The results demonstrate that PLDT's intermediate approach effectively combines the advantages of both discrete and continuous emotion representations. Our findings suggest that carefully considering the trade-off between these representational paradigms is crucial for accurately modeling the complex nature of art-induced emotional responses.

1 Introduction

In recent years, visual emotion recognition has gained significant attention in the field of computer vision (CV) (Alameda-Pineda et al., 2016; Chen et al., 2015; Rao et al., 2020), with applications ranging from human-computer interaction to digital art curation. While existing approaches have achieved promising results in recognizing emotions from facial expressions and natural scenes, detecting emotions elicited by paintings remains a significant challenge due to the abstract nature of artistic expression and the inherent subjectivity of emotional responses (Achlioptas et al., 2021; Bose et al., 2021). Traditional approaches focusing on mapping visual features to discrete emotion categories prove inadequate when handling the complex emotional responses evoked by artwork. The challenge of emotion recognition in paintings stems from three key factors: the gap between visual features and subjective responses, the diversity of individual interpretations, and the lack of robust methods for aggregating multiple emotional perspectives. These challenges necessitate a novel approach that can capture both dominant emotions and subtle nuances while preserving the richness of human emotional responses.

1.1 Representation of Emotional Responses

One-hot encoding, the conventional approach to emotion classification, fails to capture the nuanced interplay of multiple emotions that viewers often experience simultaneously when engaging with artwork (Bradley and Lang, 2007; Calvo and Lang, 2004). To address this limitation, we propose a comprehensive framework that bridges discrete and continuous emotion representations through Label Distribution Learning (LDL) (Geng, 2016). Our novel Progressive Label Discretization Technique (PLDT) enables flexible transition between these representations, effectively capturing both dominant emotions and subtle emotional nuances. For rigorous evaluation of this complex emotion modeling task, we propose TESA (Thresholded Emotion Set Accuracy), a novel evaluation framework that employs adaptive thresholds. This framework enables comprehensive assessment of how different methods balance between prediction confidence and emotional multiplicity, providing deeper insights than traditional rank-based metrics. The key contributions of this paper are:

- A novel emotion representation framework (PLDT) that bridges discrete and continuous approaches
- TESA, a threshold-based evaluation metric for multi-emotion prediction assessment

- Comprehensive analysis of representation methods' effectiveness in emotion modeling

Through these contributions, we establish a foundation for more accurate and nuanced emotion recognition in artistic contexts, while maintaining scientific rigor in evaluation and analysis.

2 Related Work

The field of visual emotion understanding has been studied for a long time, with emotion classification being particularly well-known(Cen et al., 2024; Xu et al., 2022; Chen et al., 2014). Traditionally, the domain of visual emotion understanding has focused on real-world photographs, such as human faces(Li and Deng, 2020). However, in recent years, more abstract domains that involve subjectivity, such as artworks and advertisements, have gained attention(Hussain et al., 2017; Aslan et al., 2022). These studies, in particular, emphasize the interpretation of emotion class prediction from images(Achlioptas et al., 2021; Aslan et al., 2022). Additionally, emotional image captioning (EIC) has garnered interest(Li et al., 2021; Zhao et al., 2020; Wu and Li, 2023). EIC models aim to describe visual content with emotional words (e.g., "beautiful" or "lonely"), enhancing the appeal and uniqueness of textual descriptions.

To overcome the limitations of one-hot encoding, researchers have proposed various approaches, with label smoothing(Szegedy et al., 2016; Pereyra et al., 2017) and Label Distribution Learning (LDL) being particularly noteworthy. Label smoothing is a simple yet effective technique to prevent model overfitting and adjust the confidence of predictions. This technique smooths the one-hot encoding by adding a small probability value to the correct label.

On the other hand, Label Distribution Learning (LDL) is a more direct approach to handling label ambiguity. In LDL, labels are represented as a discrete probability distribution for each sample. The core idea is for the model to predict the entire distribution of labels rather than a single class. During the learning process, LDL minimizes the distance between the actual label distribution and the predicted distribution by the model. Typically, KL divergence(Kullback and Leibler, 1951)is used as the distance metric:

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (1)$$

where P is the actual label distribution and Q is the predicted distribution by the model.

3 Proposed Method

We propose a methodology for evaluating emotional understanding of artworks by visual models, focusing on the aggregation and analysis of human emotional responses. We examine three candidate approaches for emotional label representation, spanning from discrete to continuous representations.

3.1 Emotional Opinion Aggregation

We introduce a novel approach for aggregating emotional opinions to assess how well visual models interpret emotional responses to artworks. This method consolidates subjective emotional evaluations from multiple annotators into a unified emotional representation.

3.1.1 Integration of Emotional Evaluations

Required Data

1. **Emotion Categories**: Define a set of emotion categories $E = \{e_1, e_2, \dots, e_k\}$, where k denotes the number of distinct emotional categories.

2. **Annotators**: Define a set of annotators $A = \{a_1, a_2, \dots, a_n\}$, where n is the total number of annotators involved.

3. **Emotion Evaluation Vectors**: Each annotator a_i provides an evaluation vector $v_i = [v_{i1}, v_{i2}, \dots, v_{ik}]$, where v_{ij} represents the evaluation score assigned by a_i to the emotion category e_j.

Aggregation of Evaluation Scores

To synthesize the evaluations across all annotators for each emotion category, the following aggregation is performed:

$$s_j = \sum_{i=1}^{n} v_{ij}, \quad j = 1, 2, \dots, k$$

where s_j represents the aggregated evaluation score for the emotion category e_j.

Normalization

The aggregated scores are normalized to produce a probability distribution across the emotion categories:

$$p_j = \frac{s_j}{\sum_{l=1}^{k} s_l}, \quad j = 1, 2, \dots, k$$

Here, p_j denotes the normalized probability for emotion category e_j.

3.1.2 Representation of Emotional Probability Distribution

The final emotional representation is expressed as a probability distribution $p = [p_1, p_2, \ldots, p_k]$, where:

- p_j represents the probability associated with emotion category e_j.
- The probabilities sum to one: $\sum_{j=1}^{k} p_j = 1$.
- Each probability value satisfies $0 \leq p_j \leq 1$.

This approach ensures that subjective evaluations from multiple annotators are effectively consolidated into a comprehensive emotional representation. The resulting probability distribution captures the collective emotional response elicited by the artwork, facilitating both general classification and refined distribution calibration for advanced classification models.

3.2 supervisory signal representations

Several candidate methods can be considered for representing the teacher signal, including the method we propose. In this section, we will explain the definition of each.

3.2.1 One-hot Encoding

As a baseline for the teacher signal, we adopt One-hot Encoding, which is widely used in class classification problems. This method converts categorical variables into numerical vectors, where each emotion is represented as a binary vector with a single "1" indicating the presence of that emotion. In the context of emotion classification, this approach assumes that each artwork primarily evokes a single dominant emotion, providing a clear learning objective despite simplifying the complex nature of emotional responses. The mathematical representation is as follows:

$$O(c, k) = [o_1, o_2, ..., o_k], \quad \text{where} \quad o_i = \begin{cases} 1 & \text{if } i = c \\ 0 & \text{otherwise} \end{cases}$$

3.2.2 Label Distribution Learning (LDL)

In the task of aggregating emotional opinions, we apply Label Distribution Learning (LDL), utilizing the Kullback-Leibler (KL) divergence as our loss function to optimize the predicted distribution towards the true distribution. While LDL is typically employed in tasks with clear correct answers to enhance robustness(Geng et al., 2013; Xu et al., 2014; Gao et al., 2017), our application is motivated by its unique advantages in representing complex emotional signals. LDL has the potential to naturally represent coexisting emotions, express prediction uncertainty through class probabilities, and capture subtle differences between similar emotions. These capabilities are crucial in emotion prediction tasks, where emotions are often complex and multifaceted(Mohamed et al., 2022b). For example, when a painting simultaneously evokes "sadness" and "nostalgia," LDL can represent this as a probability distribution rather than forcing a binary choice. The resulting probability distributions provide insights into both the presence and intensity of different emotions, offering a richer understanding of emotional responses to artwork. This makes the output more nuanced and informative compared to traditional single-category classifications.

3.2.3 Progressive Label Distribution Transition (PLDT)

We propose the Progressive Label Distribution Transition (PLDT) as a flexible framework that enables bidirectional conversion between traditional one-hot encoding classification and Label Distribution Learning (LDL). PLDT offers two complementary approaches: PLDT-A, which transitions from distributions to one-hot labels (LDLOnehot), and PLDT-B, which progresses from one-hot labels to full distributions (OnehotLDL). This bidirectional capability allows models to adapt to different learning scenarios and requirements. Operating based on the principle of progressive adaptation(Tzeng et al., 2015; Kumar et al., 2020), PLDT can initiate training from either end of the spectrum. PLDT-B begins with distinct one-hot encoded labels and gradually introduces the complexity of full label distributions over specified epochs, helping models develop more nuanced emotional representations. Conversely, PLDT-A starts with complete label distributions and progressively sharpens them into one-hot encodings, encouraging the model to develop clearer decision boundaries. This dual approach enables models to flexibly adapt their learning strategy based on specific task requirements. For both directions, we utilize a single interpolation function that combines one-hot encoded labels and full label distributions:

$$I(h, p) = (1 - p) \cdot O(h) + p \cdot h \tag{2}$$

where h represents the input label distribution histogram, p denotes the transition progression (with $0 \le p \le 1$), and $O(h)$ is the one-hot encoding of h. The transition progression p controls the direction and degree of transformation: in PLDT-B (OnehotLDL), p increases from 0 to 1, while in PLDT-A (LDLOnehot), p decreases from 1 to 0. This unified formulation provides a smooth and controlled transition in either direction, allowing models to progressively adapt to different label representations while maintaining learning stability.

3.2.4 Selective Distribution Dampening Loss (SDDL)

We propose a novel method called *Selective Distribution Dampening Loss (SDDL)*, drawing inspiration from the concept introduced in Focal Loss of adjusting learning intensity based on the "hardness" or rarity of samples. In our approach, we aim to maintain focus on the dominant emotional signals while selectively down-weighting extremely rare opinions (probabilities). Although Label Distribution Learning (LDL) excels in representing multiple coexisting emotions, it can sometimes overemphasize minor elements in the target distribution. To address this, SDDL introduces a threshold-based weighting mechanism that modulates the contribution of each class according to its probability in the target distribution.

Formally, let t be the target distribution and $\hat{t}$ be the predicted distribution, both of which are K-dimensional probability distributions. We first compute the Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(t \,\|\, \hat{t}) = \sum_{k=1}^{K} t_k \Big[\ln\big(t_k + \epsilon\big) - \ln\big(\hat{t}_k + \epsilon\big)\Big] \tag{3}$$

where ϵ is a small constant (e.g., 1×10^{-6}) for numerical stability. Next, we introduce a threshold parameter τ (e.g., 0.3) to distinguish "important" classes ($t_k \ge \tau$) from those considered "less important" ($t_k < \tau$). We define a weighting function:

$$w_k = \begin{cases} 1 & \text{if } t_k \ge \tau \\ \left(\frac{t_k}{\tau}\right)^{\gamma} & \text{otherwise} \end{cases} \tag{4}$$

where γ controls how aggressively classes below τ are dampened. Finally, the SDDL objective is given by:

$$\mathcal{L}_{\mathrm{SDDL}} = \sum_{k=1}^{K} w_k \, t_k \Big[\ln\big(t_k + \epsilon\big) - \ln\big(\hat{t}_k + \epsilon\big)\Big] \tag{5}$$

We then sum over all classes and average across samples to obtain a differentiable loss, which shifts attention toward classes whose target probabilities exceed τ while dampening the influence of extremely rare classes $(t_k < \tau)$. Increasing γ intensifies suppression of small t_k, thereby reducing their effect on parameter updates.

This approach is particularly beneficial in situations where maintaining a distributional representation is crucial, yet overly small probabilities can destabilize training or dilute the emphasis on dominant emotional cues. By balancing continuous distribution representation and selective emphasis, SDDL complements the advantages of LDL while preventing negligible probabilities from overshadowing the primary signals.

4 Dataset

4.1 Emotion Elicitation in Painting Datasets

For tasks like opinion aggregation in this research, it is essential to have datasets where multiple annotations are made fairly for a single data point. However, such datasets are currently rare. Representative datasets for emotions elicited by paintings include ArtEmis, ArtPedia, and WikiArt Emotions(Mohammad and Kiritchenko, 2018; Stefanini et al., 2019).

Achlioptas et al. proposed the ArtEmis dataset, a large-scale dataset that links artworks to human emotions. This dataset is frequently used in research related to the arts. It primarily focuses on the emotional experiences evoked by visual artworks and includes basic information about the artworks, emotional annotations by humans, and natural language explanations for why each emotion was elicited. The dataset is built on WikiArt and covers 27 art styles (e.g., abstract, cubism, impressionism) and 45 genres (e.g., landscape, portrait, still life), including 80,031 unique works by 1,119 artists.

In the ArtEmis dataset, at least five annotators were asked to choose one emotion from the following nine categories after viewing an artwork and then explain why they chose that emotion:

amusement, awe, contentment, excitement, anger, disgust, fear, sadness, something else

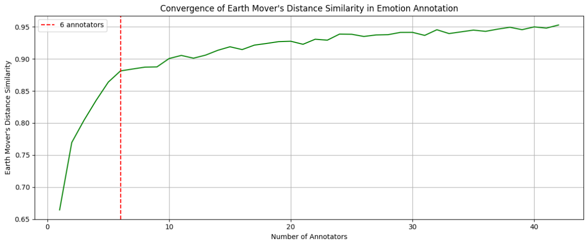

Figure 1: Graph showing reliability evaluation of emotion distribution shape reproducibility using EMD, based on the number of annotators.

This emotion model originally consisted of eight categories, but was extended by adding a ninth category, "something else," which represents either emotions not explicitly listed or the absence of a strong emotional response, such as indifference to the presented artwork. The ArtEmis dataset, which includes subjective emotional voting data from individual annotators, is well-suited as an opinion aggregation dataset.

On the other hand, datasets like ArtPedia, which includes emotional reactions to paintings along with descriptions of the painting's content and cultural background, and WikiArt Emotions, which includes emotions and art styles related to paintings, assign a single emotion label per image based on the most likely or majority emotion. These datasets are not suitable for opinion aggregation tasks since they do not collect the opinions of multiple annotators.

Therefore, a dataset like ArtEmis, which includes individual annotations from multiple annotators, is more appropriate for the tasks described in this research.

4.2 Validity as an Opinion Aggregation Set

While ArtEmis is capable of being used for opinion aggregation tasks, there are concerns regarding its reliability as a dataset for aggregated opinions. The expressive power of the opinion distribution depends on the number of annotators, and in ArtEmis, 96% of the annotations are contributed by just 5 or 6 annotators. This limited number is expected to be insufficient to represent the full spectrum of emotional opinion aggregation. Figure 1 analyzes the reliability of emotion distributions with varying annotator numbers using Earth Mover's Distance (EMD). We conduct simulations using ArtEmis samples with over 42 annotators (approximately 700 cases) as the ground truth distributions. For each annotator count (1 to 42), we perform 100 simulations of multinomial sampling and compute the normalized EMD between sampled and ground truth distributions.

The results show that distribution reliability improves significantly with increasing annotators before plateauing. While ArtEmis uses 6 annotators (red dashed line), our analysis indicates that 11 annotators are needed to achieve 95% of maximum reliability, suggesting that current ArtEmis annotations may not fully capture reliable emotion distributions.

4.3 ArtElingo

In this study, we utilize the ArtElingo dataset(Mohamed et al., 2022a), which is an extension of ArtEmis. ArtElingo includes annotations in Arabic, Chinese, and Spanish, encompassing over 51,000 images. After removing the extremely sparse annotations in Spanish, the number of annotators ranges from 5 to 76, with an average of 13.87 annotators per image. By considering English as a representative language of the West, Chinese for the East, and Arabic for the Middle East, the dataset encompasses a broad and diverse global representation. This diverse linguistic inclusion makes ArtElingo more suitable as a dataset for aggregating human emotional opinions.

5 Experiment

5.1 Overview

In this section, we train a visual model using painting images as input, with the emotion probability distributions constructed from the ArtElingo annotation data as the ground truth. We then perform a comparative analysis of the four methods presented in Section 3.2.

5.2 Data Processing

5.2.1 Image Data Processing

In this study, where we handle the delicate visual features of paintings, special care must be taken in selecting data augmentation techniques(Cetinic et al., 2018; Shorten and Khoshgoftaar, 2019). Many powerful data augmentation methods commonly used in general image classification tasks may distort the intrinsic features of paintings, making them difficult to apply. Therefore, we have carefully selected two specific augmentation methods: random cropping, which allows the model to focus on different parts of the painting during training, and random horizontal flipping, as this transformation typically does not significantly alter the overall impression of paintings.

These methods were specifically chosen to preserve critical artistic elements while providing beneficial variations for model training. They maintain the original composition, color integrity (essential for emotional expression), and textural elements such as brushstrokes, while preserving each artist's unique style. While more aggressive augmentation methods might enhance model generalization, we prioritize preserving the authentic emotional content of the artwork.

For training efficiency, all images are resized to have their shorter side set to 224 pixels while maintaining the aspect ratio, followed by random cropping to 224×224 pixels. This approach reduces computational complexity while preserving essential visual information.

5.2.2 Dataset Filtering and Splitting

Figure 2 presents a histogram of emotion labels from all annotators, revealing significant data imbalance(Achlioptas et al., 2021; Mohamed et al., 2022a) where "contentment" is the most frequent emotion and "angry" is notably rare. This imbalance is particularly pronounced when considering the Top-1 (most frequent) emotion for each image. To address this imbalance, we capped the number of samples per emotion at 2,000, specifically for cases where an emotion was the Top-1 label. As shown by the blue bars in Figure 3, some emotions (e.g., "excitement," "anger," and "something else") have fewer than 2,000 samples, resulting in a total dataset of 15,082 samples. The red bars indicate the total number of annotations per emotion, demonstrating reduced imbalance compared to Figure 2. The processed data was split into training, validation, and test sets (6:2:2 ratio), maintaining consistent Top-1 emotion proportions across all sets (7,313 training, 2,438 validation, and 2,438 testing samples).

5.3 Experimental Setup

For the model architecture in this study, we employed a fine-tuned version of the pre-trained ResNet-50 model(He et al., 2016), specifically using the Image Encoder from CLIP [34]as the base. To this, we added two fully connected layers at the final stage. The hidden layers of the added fully connected layers consist of 512 and 9 dimensions, respectively. We set the batch size to 32, and a dropout rate of 0.2 was applied. The model with the best validation loss was selected as the final model.

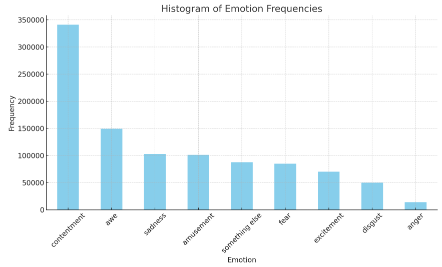

Figure 2: A histogram aggregating emotion labels provided by all annotators for each emotion.

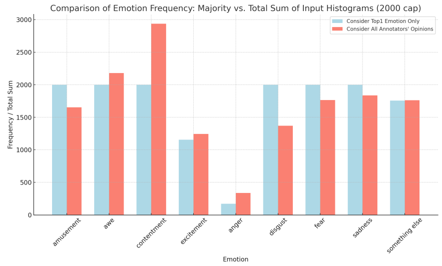

Figure 3: Comparison of Top1 sample counts (blue) and total annotation counts (red) for each emotion category. This demonstrates partial mitigation of data imbalance.

For optimization, we used the AdamW optimizer, setting the learning rate to 1e-6 for the CLIP model and 1e-4 for the added fully connected layers. A weight decay of 0.01 was applied to prevent overfitting.

5.4 Evaluation

To comprehensively evaluate the performance of emotion distribution learning models, we employ both distribution-based metrics and accuracy-based evaluation approaches. Our evaluation framework consists of distribution similarity measures and novel accuracy metrics designed specifically for multi-emotion scenarios.

5.4.1 Distribution Similarity and Rank-based Metrics

To measure the similarity between predicted and ground truth emotion distributions, we utilize Kullback-Leibler (KL) Divergence, which measures the relative entropy between predicted and ground truth distributions, while also employing rank-based accuracy metrics to assess our model's performance in identifying dominant emotions.

Specifically, we evaluate Top-1 Accuracy to measure the model's ability to correctly identify the most prominent emotion, and Top-2 Accuracy to assess the accuracy in identifying the two most prominent emotions in the correct order.

5.4.2 Thresholded Emotion Set Accuracy (TESA)

We propose a novel evaluation metric, Thresholded Emotion Set Accuracy (TESA), for assessing emotion distribution learning models. TESA enables nuanced evaluation of multi-emotion scenarios by introducing probability thresholds that determine significant emotions in both predicted and ground truth distributions. At its core, TESA computes the intersection-over-union of emotion sets that exceed a given threshold in both predicted and ground truth distributions:

$$TESA_\tau = \frac{|T(\tau) \cap P(\tau)|}{|T(\tau) \cup P(\tau)|} \tag{6}$$

where $T(\tau) = \{i : t_i \geq \tau\}$ represents the set of emotions whose true probability exceeds τ, and $P(\tau) = \{i : p_i \geq \tau\}$ represents the set of emotions whose predicted probability exceeds τ.

To provide comprehensive evaluation across different emotion multiplicities, we analyze TESA at specific thresholds τ_n where the ground truth distribution contains exactly n emotions. These thresholds are determined by:

$$\tau_n = \arg\min_\tau |E[|T(\tau)|] - n| \tag{7}$$

where $E[|T(\tau)|]$ denotes the expected number of emotions exceeding threshold τ across the dataset. Our analysis covers scenarios with varying numbers of significant emotions by evaluating $n \in \{1, 2, 3, 4\}$. For a test set with M samples, we compute the mean TESA score as:

$$\overline{TESA}_n = \frac{1}{M} \sum_{k=1}^{M} TESA_n^k \tag{8}$$

where $TESA_n^k$ represents the TESA score for the k-th sample at threshold τ_n.

The TESA framework offers several key advantages: adaptive evaluation based on emotion intensity thresholds, direct interpretation of model performance across different emotion multiplicity scenarios, robust evaluation accounting for natural variation in emotion intensity, and clear distinction between primary and secondary emotions while maintaining distributional properties. This comprehensive framework enables assessment of both distributional accuracy and practical utility of emotion distribution learning models. Unlike traditional rank-based metrics such as Top-1 and Top-2, TESA remains effective regardless of distribution shape, entropy variations, or annotator count differences by providing flexible threshold-based accuracy evaluation.

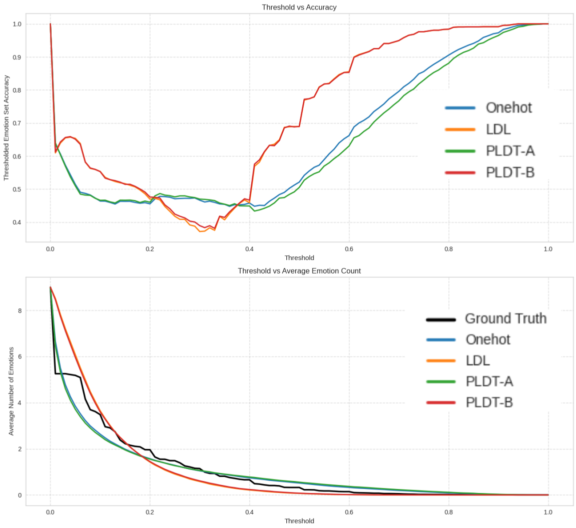

Figure 4: Analysis of threshold effects on model performance. Top: Average TESA (intersection-over-union accuracy) across test data for varying threshold values. Bottom: Average number of predicted emotions above threshold compared to ground truth distribution.

6 Results

Our experimental results demonstrate the effectiveness of different label encoding approaches

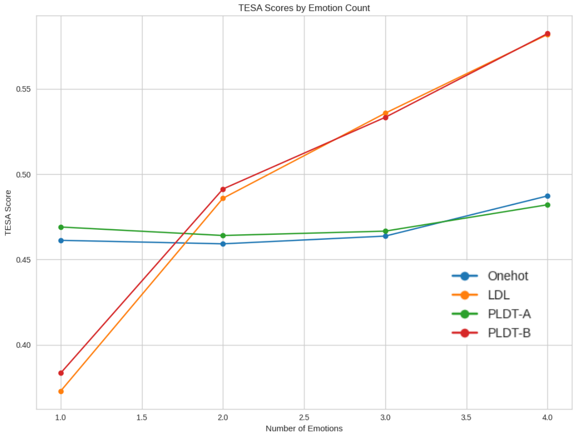

Figure 5: Comparison of Thresholded Emotion Set Accuracy (TESA) scores for different numbers of emotions (N = 1,2,3,4) across all methods.

Table 1: Comparison of One-hot, LDL, and PLDT methods across various overall performance metrics. The best score for each metric is highlighted in bold. TESA-N represents the Thresholded Emotion Set Accuracy where N indicates the target number of emotions.

Method	Basic Metrics			TESA Scores			
	KL	Top-1	Top-2	TESA-1	TESA-2	TESA-3	TESA-4
One-hot	0.719	**0.518**	0.174	0.461	0.459	0.464	0.487
LDL	**0.449**	0.503	**0.211**	0.373	0.486	**0.536**	**0.582**
PLDT-A	0.738	0.517	0.180	**0.469**	0.464	0.467	0.482
PLDT-B	**0.449**	0.497	0.206	0.384	**0.491**	0.533	**0.582**

across various metrics, as shown in Table 1. The LDL and PLDT methods show distinct advantages in different evaluation scenarios.

In terms of basic metrics, LDL and PLDT-B achieve the best KL divergence, indicating their superior ability to model emotion distribution patterns. While the One-hot method shows the highest Top-1 accuracy, LDL achieves the best Top-2 accuracy, suggesting its effectiveness in capturing multiple emotions. Notably, we observe similar performance patterns between One-hot/PLDT-A and LDL/PLDT-B pairs, indicating that the final training phase significantly influences the model's behavior.

The TESA scores reveal distinct patterns across different emotion count settings. As shown in Figure 5, PLDT-A performs best for single emotion prediction, while PLDT-B excels in dual emotion scenarios. For higher emotion counts, LDL and PLDT-B demonstrate superior performance, both achieving the highest TESA-4 scores.

Figure 4 provides insights into threshold sensitivity and its relationship with prediction accuracy. LDL and PLDT-B maintain stable performance across different threshold values, particularly at small thresholds. The emotion count analysis reveals that these methods also better align with the ground truth distribution, suggesting that accurate emotion count prediction contributes to higher TESA scores. One-hot and PLDT-A show advantages at moderate thresholds where the average emotion count approaches one, but their performance decreases at higher thresholds due to overprediction of high probability values.

An interesting phenomenon emerges in the comparison between LDL and PLDT-B: while LDL performs better in Top-k metrics, PLDT-B shows superior performance in several TESA metrics. This reversal can be attributed to their different approaches to probability distribution learning and the inherent characteristics of each evaluation metric. Top-k metrics evaluate strict ranking performance, where LDL excels due to its direct optimization of complete probability distributions, enabling precise modeling of relative emotion intensities. This advantage stems from LDL's training objective that simultaneously considers the entire probability space, leading to more accurate preservation of emotion intensity ordering.

In contrast, TESA measures the intersection-over-union of emotions above specific thresholds, where PLDT-B demonstrates superior performance. This advantage can be attributed to two key factors: First, PLDT-B's progressive transition from One-hot encoding helps maintain clearer decision boundaries for emotion activation, effectively learning appropriate threshold levels for each emotion. Second, the gradual incorporation of distribution information during training allows PLDT-B to balance between discrete and continuous representations, resulting in more robust probability estimates around decision thresholds. This unique characteristic makes PLDT-B particularly effective in scenarios where the identification of present emotions is more crucial than their exact intensity ordering.

7 Conclusion

In this work, we addressed the challenge of modeling emotional responses to artwork by exploring the spectrum between discrete and continuous label representations. Our analysis reveals that while One-hot encoding excels at identifying dominant emotions, LDL better captures subtle emotional nuances. To bridge this gap, we introduced PLDT, demonstrating that a gradual transition between these approaches can effectively balance their respective strengths. The threshold-based evaluation through TESA provided key insights into how different methods handle the trade-off between prediction confidence and emotion multiplicity. Our findings suggest that considering emotions as distributions rather than discrete labels better aligns with the complex nature of human emotional responses to art.

References

Panos Achlioptas, Maks Ovsjanikov, Kostiantyn Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11569–11579.

Xavier Alameda-Pineda, Elisa Ricci, Yan Yan, and Nicu Sebe. 2016. Recognizing emotions from abstract paintings using non-linear matrix completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5240–5248.

Sinem Aslan, Giovanna Castellano, Vincenzo Digeno, Giuseppe Migailo, Raffaele Scaringi, and Gennaro Vessio. 2022. Recognizing the emotions evoked by artworks through visual features and knowledge graph-embeddings. In *Image Analysis and Processing – ICIAP 2022 Workshops*, pages 129–140. Springer, Berlin, Germany.

Debayan Bose, Krishna Somandepalli, Sagnik Kundu, Ritam Lahiri, Jonathan Gratch, and Shrikanth Narayanan. 2021. Understanding of emotion perception from art. *arXiv preprint arXiv:2110.06486*.

Margaret M. Bradley and Peter J. Lang. 2007. The international affective picture system (iaps) in the study of emotion and attention. In James A. Coan and John J. B. Allen, editors, *Handbook of Emotion Elicitation and Assessment*, pages 29–46. Oxford University Press.

Manuel G. Calvo and Peter J. Lang. 2004. Gaze patterns when looking at emotional pictures: Motivationally biased attention. *Motivation and Emotion*, 28(3):221–243.

Jian Cen, Chaoqing Qing, Hong Ou, Xinyu Xu, and Jian Tan. 2024. Masanet: Multi-aspect semantic auxiliary network for visual sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–12.

Ela Cetinic, Tomislav Lipic, and Sonja Grgic. 2018. Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications*, 114:107–118.

Ming Chen, Lu Zhang, and Jan P. Allebach. 2015. Learning deep features for image emotion classification. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 4491–4495.

Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838.

Xiaojun Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Xiaojun Geng, Chao Yin, and Zhi-Hua Zhou. 2013. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(10):2401–2412.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1705–1715.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Ananya Kumar, Tengyu Ma, and Percy Liang. 2020. Understanding self-training for gradual domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5468–5479.

Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215.

Tao Li, Yifan Hu, and Xueming Wu. 2021. Image captioning with inherent sentiment. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Youssef Mohamed, Mohamed Abdelfattah, Sara Alhuwaider, Fei Li, Xinyue Zhang, Kenneth W. Church, and Mohamed Elhoseiny. 2022a. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Youssef Mohamed, Fatima F. Khan, Khayrullo Haydarov, and Mohamed Elhoseiny. 2022b. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21263–21272.

Saif M. Mohammad and Svetlana Kiritchenko. 2018. Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1225–1232.

Gabriel Pereyra, George Tucker, Jan Chorowski, ukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Tianyi Rao, Xiaohui Li, and Mingliang Xu. 2020. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, 51:2043–2061.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, pages 729–740.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4068–4076.

Xueming Wu and Tao Li. 2023. Sentimental visual captioning using multimodal transformer. *International Journal of Computer Vision*, 131(4):1073–1090.

Liang Xu, Zhaowei Wang, Bo Wu, and Shing-Chi Cheong Lui. 2022. Mdan: Multi-level dependent attention network for visual emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9479–9488.

Yan Xu, Tao Mo, Qiwei Feng, Eric I-Chao Chang, Yan Xu, and Lian-Ming Du. 2014. Deep learning of feature representation with multiple instance learning for medical image analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1005–1009.

Weixiang Zhao, Xueming Wu, and Xian Zhang. 2020. Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12984–12992.