A Multi-Stage Visual Perception Approach for Image Emotion Analysis

Shangfei Wang, Jicai Pan, and Jinqiao Lu

Most current methods for image emotion analysis suffer from the affective gap, in which features directly extracted from images are supervised by a single emotional label, which may not align with users' perceived emotions. To effectively address this limitation, this paper introduces a novel multi-stage perception approach inspired by the human staged emotion perception process. The proposed approach comprises three perception modules: entity perception, attribute perception, and emotion perception. The entity perception module identifies entities in images, while the attribute perception module captures the attribute content associated with each entity. Finally, the emotion perception module combines entity and attribute information to extract emotion features. Pseudo-labels of entities and attributes are generated through image segmentation and vision-language models to provide auxiliary guidance for network learning. A progressive understanding of entities and attributes allows the network to hierarchically extract semantic-level features for emotion analysis. Comprehensive experiments on image emotion classification, regression, and distribution learning demonstrate the superior performance of our multi-stage perception network.

Image emotion analysis is concerned with the recognition of emotions experienced when an image is viewed. Given the growing trend of opinion and emotion expression via images on the Internet, image emotion analysis has gained significance in the field of computer vision. Advancements in image emotion analysis can positively impact several applications, including opinion mining (Hassan et al. 2019), business intelligence (Hosany and Prayag 2013), psychological well-being (Guntuku et al. 2019), and entertainment assistance (Xing et al. 2015).

One significant challenge to accurate image emotion analysis is the affective gap, which refers to the disparity between visual features and the perceived emotions by viewers (Zhao et al. 2021). Researchers have tried to overcome this challenge by extracting discriminative features capable of better distinguishing emotional content in images. These can be either hand-crafted features or deep features. Hand-crafted feature design necessitates expertise in psychological and artistic principles, such as color, texture (Machajdik and Hanbury 2010), and scene attributes (Patterson and Hays 2012). These features may not encompass sufficient information to achieve effective image emotion analysis performance. Deep features are primarily obtained through convolutional neural networks (CNNs) (Yang et al. 2018) and attention-based networks (Zhao et al. 2019). Although deep networks perform better than methods using hand-crafted features, a substantial gap still exists between pixel-level signals and semantic-level emotional labels. They rely on visual features directly supervised by emotional labels, which struggle to capture abstract emotional information.

Most methods in emotion analysis extract features directly from images. However, Ortony, Clore, and Collins (1988) propose a process involving entity perception and evaluation. Psychological research has established the significance of image entities (Brosch, Pourtois, and Sander 2010) and the relationships between them (Bar 2004) for emotion evocation. Ortony et al. (1988) propose a cognitive model that describes the experience of emotions through a sequence of stages. First, the image is perceived and the entities within it are recognized. Subsequently, the perceived entities are evaluated based on personal experiences. Perception and evaluation culminate in the experience of a specific emotion. Ortony’s theory emphasizes visual contents like entities and attributes within the image, highlighting their role in the perception and evaluation processes. As illustrated in Figure [1](#fig:example), an image comprised of a joyful woman, beautiful flowers, and a spectacular mountain will evoke joy in the viewer. Rather than directly extracting features from the image, Ortony’s theory aligns with the human emotion generation process.

We drew inspiration for Ortony’s theory to address the challenge posed by the affective gap. Our multi-stage visual perception network for image emotion analysis includes attention-based modules dedicated to extracting entity, attribute, and emotion features, as depicted in Figure [[fig:framework]](#fig:framework). The entity module combines deep feature maps from multiple levels for each entity, employing entity pooling and an attention mechanism. The fused entity features, along with the high-level deep feature map, are fed into the attribute module, which derives the attribute feature for each entity through entity pooling. The emotion module integrates all entity and attribute features into an emotion feature via global pooling and attention operations. To address the lack of entity and attribute labels in most image emotion data sets, we generate pseudo labels for entities and attributes. Ultimately, the proposed network is optimized by leveraging entity, attribute, and emotion labels. To validate the effectiveness of our approach, we conduct comprehensive experiments on eight data sets. The contributions of this paper are summarized as follows:

First, we propose a multi-stage visual perception approach inspired by the human emotion perception process, bridging the affective gap in a staged manner. Secondly, we design entity, attribute, and emotion modules based on attention mechanisms to extract multi-level features. Thirdly, we conduct experiments showing the superior performance of the proposed network on various image emotion analysis tasks.

# Related Work

Image emotion analysis can be divided into three main tasks: image primary emotion classification, image emotion regression, and image emotion distribution learning. Here, we provide a brief introduction to the research status of these three tasks. For more detailed information, refer to the comprehensive review paper by Zhao et al. (Zhao et al. 2021).

## Emotion Representation

Research has found two main types of emotion representation models in the field of affective computing: categorical emotion and dimensional emotion space (Zhao et al. 2021). Categorical emotion models classify emotions into several basic categories. Common categories include Ekman’s six basic emotions (Ekman 1992), happiness, sadness, fear, surprise, disgust, and anger; Plutchik’s eight basic emotions (Plutchik 1984), joy, sadness, disgust, surprise, fear, anticipation, trust, and anger; and Mikel’s eight basic emotions (Mikels et al. 2005), joy, sadness, fear, astonishment, disgust, anger, excitement, and satisfaction. Emotions can also be categorized as positive, negative, or neutral. Dimensional emotion space models typically represent emotions using continuous points in a two-dimensional or three-dimensional Cartesian space. Valence-arousal-dominance (VAD) (Schlosberg 1954) is the most widely used dimensional emotion space, where valence represents the pleasantness of the emotion ranging from negative to positive, arousal represents the intensity of emotions from calm to excitement, and dominance represents the degree of control over the emotion. Categorical emotion models are discrete labels primarily used for emotion classification tasks. For emotion distribution learning, each image is annotated using the categorical emotion model, assigning a score to each emotional category. These scores are then normalized to form an emotion distribution label. Dimensional emotion space models represent emotions as multidimensional continuous values and are chiefly used for emotion regression tasks. Dimensional emotion space can also discretize the continuous values of each dimension into several constant values for emotion classification tasks.

## Image Emotion Classification

Image emotion classification is usually a single-label classification task wherein the emotion is represented by categorical emotion states. Early studies mainly employ machine learning classifiers to predict emotion based on hand-crafted features. Machajdik et al. (Machajdik and Hanbury 2010) extracted a set of features based on psychology and art theory (e.g., color, texture, composition, and content), and used naive Bayes as the classifier. Zhao et al. (Zhao et al. 2014) extracted and combined artistic principles like balance and harmony to classify the image emotion. Due to design difficulties and limited information, these hand-crafted feature-based methods have recently been replaced by deep learning methods. Xu et al. (Xu et al. 2014) directly extracted deep features via a pre-trained CNN-based model, and trained a logistic regression model above these features. This work improved the performance of image emotion recognition by a large margin, resulting in explosive development of deep learning techniques for image emotion classification. Zhu et al. (Zhu et al. 2017) utilized a bidirectional recurrent neural network to integrate features from different CNN layers, exploiting the dependency among different levels of features. Considering that different regions have a different influence on the intended expression, Yang et al. (Yang et al. 2018) and Zhang et al. (Zhang and Xu 2020) learned the emotion intensity map via a cross-spatial pooling strategy and class activation mapping technique, respectively. Each then combined the intensity map and holistic CNN-based feature map for emotion classification. Zhao et al. (Zhao et al. 2019) employed the attention mechanism on spatial and channel dimensions of CNN-based feature maps to adaptively highlight the important regions and channels. Xu et al. (Xu et al. 2022) utilized the feature pyramid network and attention mechanism to learn multi-level emotion features based on three-level emotion labels. Since regional interactions play an important role in image semantic understanding, some researchers introduced region correlation learning. Wu et al. (Wu et al. 2021) and Yang et al. (J. Yang, Gao, et al. 2021) employed graph convolutional networks to enhance emotion features by mining relationships between distinct objects. Zhang et al. (Zhang et al. 2022) employed Transformer to learn the associations between different regions for emotion recognition. Over time, feature extractors have become increasingly complex. Despite their complexity, they are limited by the affective gap as features are directly extracted from images with the supervision of emotion labels only.

Previous works have explored the use of auxiliary guidance for image emotion classification. For instance, Borth et al. (Borth, Chen, et al. 2013) introduced tagged images using the concept of Adjective-Noun Pair (ANP), which consists of an adjective indicating emotions and a noun corresponding to objects or scenes. Subsequently, Chen et al. (Chen et al. 2014) trained deep CNN models to classify ANP image tags and leverage them for image emotion classification. Deng et al. (Deng, Wu, Shi, Xing, Hu, et al. 2022) incorporated entities and emotional labels into a sentence and fed both image and sentence into the CLIP model (Radford et al. 2021) for feature learning. These methods utilize fine-grained labels such as ANPs and sentences to provide precise guidance for model learning, thereby enhancing the feature’s capability. However, large amounts of learning data are necessary to directly align visual features and fine-grained labels.

In summary, most existing methods directly map visual features to emotional or auxiliary labels, which does not effectively bridge the affective gap. In contrast, this paper presents a multi-stage approach that divides the affective gap into three parts with the help of emotion and auxiliary labels, thus reducing learning difficulty and enhancing features.

## Image Emotion Regression

Image emotion regression, which predicts dimensional emotions, presents a greater challenge than classification. Images belonging to the same emotion category may differ in continuous space, like valence-arousal (Schlosberg 1954). A common practice is to transfer the image emotion classification model to regression by replacing the classifier with a regression layer, such as the emotion classification models SentiBank (Borth, Ji, et al. 2013), ResNet-101 (He et al. 2016), FT ResNet-101 (You et al. 2016), SentiNet-A (Song et al. 2018), and WSCNet (Yang et al. 2018). However, each image must be mapped to a specific value in a high-dimensional continuous space, so the expansion of the prediction space makes it difficult to distinguish the features extracted by these deep models. Zhao et al. (Zhao et al. 2019) proposed the deep model PDANet for image emotion regression. They applied the attention mechanism on the spatial and channel-wise levels of the deep CNN feature maps to learn the emotional feature. PDANet has performed well on emotion regression compared to deep networks for emotion regression. The emotion regression task requires more refined image features, thus exhibiting better discriminative power in high-dimensional continuous emotional space. However, directly mapping image features to continuous emotional scores does not circumvent the affective gap. The proposed multi-stage perception network employs attention mechanisms at each stage to extract fine-grained features, providing a more powerful feature representation for addressing the affective gap in the continuous emotional space.

## Image Emotion Distribution Learning

Existing works on image emotion classification or regression typically treat the exploration of emotions in images as a single-label prediction task. However, this approach struggles to accurately capture the nuances and utility of each individual label. To overcome this, several methods have turned to label distribution learning (LDL) to describe the emotions associated with an image. Strategies employed for LDL can be categorized into four groups: problem transformation, algorithm adaptation, specialized algorithms (Geng 2016), and deep models. The first strategy, problem transformation, involves converting the LDL problem into existing learning paradigms. For instance, the problem transformation support vector machine (PT-SVM) and problem transformation naive Bayes (PT-Bayes) utilize SVM and naive Bayes, respectively, to predict the emotion labels. The second strategy, algorithm adaptation, aims to adapt existing algorithms to address the specific problem. Examples include the algorithm adaptation k-nearest neighbor (AA-kNN) and algorithm adaptation back propagation network (AA-BP). The third strategy designs specialized algorithms for label distribution learning. Geng et al. (Geng, Yin, and Zhou 2013) proposed the specialized algorithm improved iterative scaling (SA-IIS) and specialized algorithm quasi-Newton methods (SA-BFGS) for label distribution learning. The fourth strategy involves deep learning.

Deep learning has demonstrated remarkable performance in the field of computer vision, leading to the adoption of deep models for emotion distribution learning. Building upon this, Peng et al. (Peng et al. 2015) introduced a convolutional neural network regression model (CNNR) that extends the CNN architecture by replacing the softmax layer with the Euclidean loss layer to facilitate emotion distribution regression. Yang et al. (Yang, She, and Sun 2017) proposed a multi-task deep framework that jointly optimizes classification and distribution prediction. Xiong et al. (Xiong et al. 2019) developed a structured and sparse annotations framework to fuse the polarity and intensity of emotions into a CNN-based model. Xu et al. (Xu and Wang 2021) used an attention mechanism to extract emotion-related regions for each emotion, and captured the correlations between emotions using a graph-based network. Yang et al. (J. Yang, Li, et al. 2021) designed a circular representation of emotion distribution and trained the deep model with progressive circular loss. Yang et al. (Yang et al. 2022) built a subjectivity-appraising module consisting of an attention-based memory to simulate the process of human emotion perception.

Overall, these methods of image emotion distribution learning focus on capturing the correlations between emotions, but they overlook the challenge of bridging the affective gap between features and emotion labels. In contrast, the proposed multi-stage perception network introduces multiple-attribute perception stages, which can simultaneously learn emotion correlations and narrow the affective gap.

# Problem Statement

Let be an image emotion data set, where and represent image data and emotion labels, respectively, and is the total number of samples. The proposed method utilizes image entity and attribute information to learn visual features in a staged manner. Therefore, entity and attribute pseudo-labels are generated for the image emotion data set during the pre-training phase of the feature extraction network. The entity label consists of predicted results of image instance segmentation , where and are binary masks and entity category labels for the -th entity in the -th image, and is the number of entities in the -th image. The attribute content is represented by the distribution of fixed attribute words , where denotes the matching score between the -th attribute word and the -th image. With the help of entity and attribute information, three perception tasks are conducted to optimize the feature extraction network. The entity classification task trains the entity perception module on the data . The attribute learning task optimizes the attribute perception module on the data . The emotion analysis task trains the emotion perception module on the standard image emotion data set . Ultimately, these three tasks facilitate the extraction of emotion features based on the three-stage visual perception framework.

# Methodology

## Label Generation

A three-step framework is used to create the entity and attribute pseudo labels needed during the learning phase. Entity masks and categories are generated via Maskformer, an image segmentation model created by Cheng, Schwing, and Kirillov (2021).

We chose to annotate the attribute distributions by entity instead of by category, as evaluations are subjective. The CLIP (Radford et al. 2021) model is the basis of our method for annotating attribute distributions. We pulled 66 emotion-based adjectives from Plutchik’s wheel of emotion (Plutchik, 1984) to create an attribute dictionary (see Figure [2](#fig:lb)). Inspired by research on generating prompts (Deng, Wu, Shi, Xing, Hu, et al. 2022), we input the template “The [entity] in this image is [attribute]” into CLIP (see Figure [3](#fig:la)). Visually, the input consists of the entity region cropped from the original according to the entity mask. With these inputs, the CLIP model yields similarities between the entity and each of the attributes. A softmax operation converts this to an attribute distribution, denoted as , in which represents the attribute distribution of any given image scene.

## The Proposed Network

Our multi-stage perception network is made up of the backbone and the entity, attribute, and emotion modules (see Figure [[fig:framework]](#fig:framework)). The backbone extracts four levels of deep visual feature maps ( to leverage ample visual features. The entity module uses pooling and attention operations to create an entity feature. Next, that entity feature is input into the attribute model, along with the top-level feature map from the backbone. This yields each entity’s attribute feature. The attribute feature and global visual feature are integrated via the emotion module, obtaining the emotion feature used to analyze image emotion.

For each input image, an image segmentation model is used to find entity masks , where represents the number of entities in the image. Entity masks are made up of background and entity values, shown as 0 and 1 respectively. These masks are used for the entity feature pooling operation, defined as:

in which represent the feature map width and height.

### Entity Module

The human eye is first drawn to an entity in an image. This makes the entity essential for emotion analysis. Visual features for each entity are extracted via the following method. For the -th entity in an image, each level feature map is pooled into a vector based on the entity mask, as shown below:

in which is the parameter matrix of the -th level used to map varying level features onto a single common space. This formula makes the assumption that the mask is scaled to the same proportions as each level feature map.

Level vector weights are calculated using an attention mechanism as follows:

in which the parameter matrix maps each of the vectors into a scalar. A softmax operation converts the scalars to normalized weight vectors demonstrating the priority of each level feature. Lastly, we calculate the entity feature of the -th entity as:

in which and is the number of entities in the given image.

### Attribute Module

The attribute module is used to extract features representing the emotional characteristics of the entities. It bridges the gap between the entities themselves and the emotions those entities evoke.

Regions of an entity differ in importance when evaluating attributes. For example, a person’s facial region would be more relevant to determining the attribute than the rest of the body. Given that, an attention mechanism is used to determine the weight of each section of the entity as follows:

in which parameter matrix maps each location feature into a scalar. Feature map possesses higher-level semantic information that can be used to extract the features of the attributes. The value of is set at negative infinity, so the background information is given little weight.

The attribute module can also appraise the image scene, providing insight into the overall emotion. Based on the global image, the scene weight can be calculated as follows:

The final attribute output model combines the weighted and identity features thusly:

in which and . Entity information for the entire image is represented by .

### Emotion Module

The emotion module integrates entity and attribute features for overall emotion analysis. The following attention-based global pooling formula is used to extract global feature :

in which each location feature of the image is mapped into a scalar via parameter matrix .

Entities within an image vary in how much they contribute to the overall emotion. To accommodate this, an attention mechanism calculates the weights of the entities as follows:

in which is the parameter matrix.

Lastly, the emotion feature is created by combining the global feature and the weighted information:

In its entirety, the emotion feature encompasses entity, attribute, and global semantic information.

## Loss Function

Entity classification, attribution distribution, and emotional analysis are the three tasks we study for model learning. Cross-entropy loss of entity classification is computed during the entity perception stage with the following formula:

in which represents the entity category.

The distribution of emotional attributes is highly nuanced, introducing subjectivity into the attribute perception task. Kullback-Leibler loss, which measures the distance between predicted and labeled distributions, is used to optimize this module:

in which is the number of attributes and is the probability of the -th attribute.

Different loss functions are used for various tasks during the emotion perception phase. Emotion classification employs cross-entropy loss:

in which is the ground truth label for the image.

Emotion regression utilizes mean square error loss:

in which is the number of emotional space dimensions.

As in the attribute perception task, Kullback-Leibler loss is used during emotion distribution learning as follows:

in which represents the number of emotional categories.

The whole network is optimized by reducing total loss overall as follows:

in which stands for the penalty coefficient.

# Experiments

## Data sets

We conduct extensive experiments on eight data sets, including the emotion classification data sets EmotionRoI (Peng et al. 2015), FI (You et al. 2016), and WEBEmo (Panda et al. 2018); the emotion regression data sets IAPS (Lang et al. 1997), NAPS (Marchewka et al. 2014), and EMOTIC (Kosti et al. 2017), and the emotion distribution learning data sets Flickr\_LDL(Yang, Sun, and Sun 2017) and Twitter\_LDL(Yang, Sun, and Sun 2017).

The EmotionRoI (Peng et al. 2015) data set was collected by searching the Flickr platform using Ekman’s six basic emotions (joy, sadness, fear, surprise, disgust, and anger) and their synonyms as keywords. A total of 1980 images were collected, with 330 images for each emotion category.

The FI (You et al. 2016) data set was collected by searching for Mikel’s eight basic emotions (joy, sadness, fear, surprise, disgust, anger, excitement, and contentment) as keywords on the web platforms Flickr and Instagram, resulting in millions of images. After removing some noisy data, 225 annotators annotated images with the eight basic emotions. Eventually, 23,308 images with manually annotated emotions were obtained.

The WEBEmo (Panda et al. 2018) data set is a large-scale web image emotion data set based on the Parrott emotional hierarchy model. The creators first searched for each emotion keyword on the internet platform, collecting about 300,000 weakly labeled images. They then removed duplicate images and images with non-English tags, resulting in 268,000 images. Each has a weak three-level emotion label with 25, 6, and 2 classes. The data set has not been manually annotated, so there may be some bias in the emotion labels.

The IAPS (Lang et al. 1997) data set is a collection of 1182 emotion-arousing natural images in photographic style. The images encompass diverse content, including portraits, babies, animals, and landscapes, among others. The data set was annotated by approximately 100 university students in the VAD dimensional emotion space. The average and variance of the annotations were taken as the emotional labels for the images. This data set has been widely used for image emotion regression tasks.

The NAPS (Marchewka et al. 2014) data set comprises 1356 images in five main classes: people, faces, animals, objects, and landscapes. The images were annotated by 204 annotators in the VAD dimensional emotion space, with an average of 55 emotional annotations per image. The mean and variance of the annotations were used as the final emotional labels for the images. This data set is extensively employed for image emotion regression tasks.

The EMOTIC (Kosti et al. 2017) data set was constructed by retrieving and selecting images from the MSCOCO data set (Lin et al. 2014), the Ade20k data set (Zhou et al. 2019), and Google search engine. It contains a total of 18316 images of individuals. Amazon Mechanical Turk annotators provided discrete emotion labels consisting of 26 emotions, and continuous VAD dimensional emotion annotations. The emotions of individuals in a total of 23788 images (66% male, 34% female) were annotated.

Flickr\_LDL (Yang, Sun, and Sun 2017) and Twitter\_LDL (Yang, Sun, and Sun 2017) are two data sets commonly used in the field of image emotion distribution learning. The Flickr\_LDL data set is a re-labeled subset of the Flickr (Borth, Ji, et al. 2013) data set. It contains 11150 images collected through adjective-noun pairs. The Twitter\_LDL data set was collected by searching for Mikel’s eight basic emotions on the Twitter platform and manually filtering out duplicate images. It consists of 10045 images. Both data sets were annotated using Mikel’s eight basic emotions, and each image was annotated by approximately ten annotators. The final emotion distribution labels were determined by the majority.

## Implementation Details

The input for the proposed network is an 224 x 224 image, obtained by resizing and center cropping each original image. The dimensions of the entity features, attribute features, and emotion features are set to 512. The backbone of the proposed network consists of ResNet-50 (He et al. 2016), ResNet-101 (He et al. 2016), and ViT-B/32 (Dosovitskiy et al. 2020); each is pre-trained using the CLIP (Radford et al. 2021) method or the ImageNet (Russakovsky et al. 2015) method. For the ResNet structure, we freeze the parameters of the first three layers, while for the ViT-B/32 model, we freeze the parameters of the first nine layers. We then train the unfrozen backbone and the entity, attribute, and emotion modules using the Adam optimizer for 50 epochs, with a batch size of 64, a learning rate of 0.00001, and a weight decay of 0.001. Following Yang et al. (2018) and Xu et al. (2022), we randomly split the FI data set into 80% for training, 5% for validation, and 15% for testing. The EmotionRoI and WEBEmo data sets are publicly split into 80% and 20% for training and testing, with 5% randomly sampled from the training set for validation. According to Zhao et al. (2019), the IAPS, NAPS, and EMOTIC data sets are randomly split into 70% training, 10% validation, and 20% testing. Referring to J. Yang, Li, et al. (2021), the Flickr\_LDL and Twitter\_LDL data sets are randomly split into 75% training, 5% validation, and 20% testing. We map the six emotional categories in the EmotionRoI data set and the eight emotional categories in the FI data set to two sentiment polarity labels, following Xu et al. (2022). For emotion classification, accuracy is used as the model performance metric. For emotion regression, mean squared error () and R squared () are used as performance metrics, referring to Zhao et al. (2019). For emotion distribution learning, we follow the evaluation metrics set in previous works (Xu and Wang 2021; J. Yang, Li, et al. 2021) and adopt six label distribution evaluation measures: Chebyshev distance, Clark distance, Canberra metric, Kullback-Leibler divergence, cosine coefficient, and intersection similarity.

## Ablation Studies

To assess the efficacy of our proposed design across various tasks, we perform ablation studies on image emotion classification, regression, and label distribution learning. These studies investigate the impact of different backbones, the role played by perception modules, and the influence of varying loss penalty coefficients.

### Visual Feature Extraction

Unlike tasks in which objective information is extracted from visual signals, image emotion analysis requires an understanding of the content and the emotions it induces, making visual feature extraction even more essential. We examined pre-training methods and backbone structures to determine how each influence results. The former includes common methods used to classify images on ImageNet (Russakovsky et al. 2015) as well as the CLIP method (Radford et al. 2021) for contrastive learning. The backbone structures we examined include ResNet-50 (He et al. 2016), ResNet-101 (He et al. 2016), and ViT-B/32 (Dosovitskiy et al. 2020). Table [[tab:backbone]](#tab:backbone) shows the results of our experiments.

Typically, the best performance is achieved when models are pre-trained using CLIP rather than ImageNet. For example, the ViT-B/32 model, which uses CLIP, performs better than ImageNet models by 0.1953 on the EmotionRoI data set, 0.2212 on the FI data set, and 0.1418 on the WEBEmo data set. That model also has better image emotion regression performance, improving over others by 1.661 on the IAPS data set, 3.3675 on the NAPS data set, and 0.4204 on the EMOTIC data set. It achieves fairly comparable performance when it comes to image emotion label distribution learning, with improvements of 0.1227 and 0.1734 on the Flickr\_LDL data set and Twitter\_LDL data set, respectively. These results demonstrate that the CLIP method is superior for image emotion analysis tasks. Because CLIP is based on a vison-language model and utilizes a contrastive learning task, it has more image-text information than ImageNet provides. Therefore, the model is better able to learn representative features.

Models pre-trained on ImageNet achieve superior performance on all data sets when ResNet architecture is used in place of the visual Transformer structure. Conversely, models pre-trained with CLIP perform better when a Transformer structure is utilized, except on the EMOTIC data set. ResNet is better at learning local information and therefore performs object classification well. In contrast, the Transformer architecture excels at learning global semantic information, and is better at learning with text supervision. We used these findings to choose the ViT-B/32 model, pre-trained with the CLIP method, as the backbone for visual feature extraction.

### Perception Modules

Ablation experiments were performed on entity, attribute, and emotion modules to further examine their impact on image emotion analysis. For these experiments, the backbone and classifier alone (without any perception modules) is used as the baseline.

As Table [[tab:perception]](#tab:perception) shows, the baseline model yields emotion recognition accuracies of 0.6599 on the EmotionRoI data set, 0.7612 on the FI data set, and 0.3700 on the WEBEmo data set. The addition of the emotion perception module improves performance by 0.0573, 0.0218, and 0.0135 on the respective data sets. When the attribute perception model is added, performance is reduced slightly. The incorporation of the entity perception module makes the most drastic improvements, improving performance by 0.0674, 0.0309, and 0.0133 on the three respective data sets. The regression and label distribution tasks of each data set mimic these findings. The multi-stage model improves by 0.4341, 0.8642, and 0.1610 on the IAPS, NAPS, and EMOTIC data sets, and by 0.0993 on the Flickr\_LDL and 0.1452 on the Twitter\_LDL data sets. For the IAPS data set and EMOTIC data set, the best performance was achieved when only the emotion module was added. This could be due to the fact that these data sets are predominantly landscapes and portraits, which tend to have fewer entities and less attribute information. Thus, the addition of the entity and attribute models had little impact on performance. In general, the emotion perception module yielded the most gains in performance overall. When the attribute perception module was added by itself, model performance decreased slightly. This could be because the affective gap is larger when attribute distribution is predicted directly from visual features. The simultaneous addition of the entity perception module divides the affective gap between entity and attribute feature extraction, making it easier to learn and analyze emotion. The results affirm the rationale for three-stage perception modules for image emotion analysis. The modules progressively learn emotion features based on an understanding of the visual content by identifying entities, evaluating their attributes, and providing additional emotional cues.

### Loss Penalty Coefficients

The influence of entity, attribute, and emotion tasks are represented by loss penalty coefficients , , and , respectively. Table [[tab:trains]](#tab:trains) shows the effect of these tasks for several common coefficient settings. The following observations can be made.

When the emotion classification task is used by itself, it yields accuracies of 0.7239 on the EmotionRoI data set, 0.7840 on the FI data set, and 0.3829 on the WEBEmo data set. If either the entity classification or attribute distribution task is used on its own (::=0:1:1 or 1:0:1, respectively), performance declines on every data set. It’s possible that the weaker connections between entity and attribute features and emotion analysis causes a deficit in feature dependency learning. A setting of 1:1:2, which uses all three of the perception tasks, yields a significant decline in performance. Optimum results are achieved when the entity and attribute task coefficients are reduced to 0.1 and 0.8, respectively, while the emotion task coefficient remains at 2. A similar ratio yields the best results for image regression and label distribution learning, although the IAPS, EMOTIC, and Flickr\_LDL data sets performed best when the entity task is omitted entirely and the attribution and emotion coefficients are both set to 1. Again, this could be caused by the plethora of landscapes and human imagery in these data sets. Those types of images typically have more limited entity data, and so the entity task has little effect on performance. In general, the emotion perception task should have the highest loss penalty coefficient. Network learning is primarily about emotional perception; entity and attribute information can enhance it, but they should be less influential. Loss coefficients should gradually increase over time because the image understanding process is progressive. Features extracted from one task are used in the subsequent task.

## Comparisons to Related Works

We compared the performance of the proposed method with current state-of-the-art methods on image emotion classification, regression, and distribution learning tasks.

### Image Emotion Classification

Other image emotion classification techniques include SentiBank, a hand-crafted, feature-based method (Borth, Ji, et al. 2013); DeepSenBank (Chen et al. 2014), AlexNet (Krizhevsky, Sutskever, and Hinton 2017), and ResNet101 (He et al. 2016), which are all deep convolutional network-based methods; WSCNet (Yang et al. 2018), PDANet (Zhao et al. 2019), Zhang et al. (Zhang and Xu 2020), MDAN (Xu et al. 2022), and MSRCA (Zhang et al. 2022), which are all based on attention mechanism; SOLVER (J. Yang, Gao, et al. 2021), which is a graph convolutional network-based method; and PT-DPC (Deng, Wu, Shi, Xing, and Jian 2022) and SimEmotion (Deng, Wu, Shi, Xing, Hu, et al. 2022), which utilize CLIP pre-training. Like the latter methods, our method is also based on CLIP pre-training, although we utilize the attention mechanism as well. Table [[tab:comparison1]](#tab:comparision1) shows the results of our experiments.

First, the methods incorporating attention mechanism greatly outperform the others (SentiBank, DeepSenBank, AlexNet, and ResNet101) on emotion classification on five of the data sets. For example, MDAN is based on the ResNet101 structure, but introduces the attention mechanism. On the FI and WEBEmo data sets, accuracy increase from 0.6616 to 0.7641, and from 0.3214 to 0.3428, respectively. This clearly indicates the benefits of incorporating the attention mechanism into image emotion classification tasks. Also, the best performance on all five data sets was achieved by methods pre-trained using CLIP (i.e., PT-DPC, SimEmotion, and the proposed technique). CLIP pre-training improves image semantic understanding, as it obtains highly representable visual features through language supervision rather than emotional labels alone. The proposed three-stage perception network based on CLIP pre-training offers additional supervision for emotion analysis, so the model is better able to understand image semantics.

Secondly, the proposed method achieves superior performance on many of the data sets. For example, the SimEmotion achieves 0.0219 on the EmotionROI data set, while the proposed method achieves 0.0287. However, SimEmotion performs better on the FI data set, possibly because of the ratio of emotion categories in that set (6:1). SimEmotion is not as impacted by data imbalance because it uses the CLIP language supervision method during training, which is better able to understand the semantics. While the proposed method also uses CLIP, it simultaneously performs entity classification, attribute distribution learning, and emotion classification, changing the original CLIP parameters and making it more susceptible to data imbalance. Despite this, the proposed method achieves the second-best performance on the F1 data set. Overall, the three-state perception process with CLIP pre-training achieves significant improvements on image emotion classification.

### Image Emotion Regression

In this section, we compare the proposed method’s performance on image emotion regression to related works, including SentiBank (Borth, Ji, et al. 2013), which is based on handcrafted features; and deep convolutional networks such as ResNet-101 (He et al. 2016), FT ResNet-101 (You et al. 2016), SentiNet-A (Song et al. 2018), WSCNet (Yang et al. 2018), and PDANet (Zhao et al. 2019). For the regression tasks on the IAPS, NAPS, and EMOTIC data sets, experimental results are shown in Table [[tab:comparison2]](#tab:comparision2). The results demonstrate that our proposed method achieves the best performance on all three data sets. The average and average on the IAPS data set are 0.993 and 6.479, respectively, showing improvements of 0.900 and 2.931. On the NAPS data set, the average and average are 0.703 and 7.690, respectively, improving by 0.968 and 3.063. On the EMOTIC data set, the average and average are 2.378 and 2.269, respectively, improvements of 0.326 and 0.867 compared to the related works. Our proposed method achieves significant improvements on all three data sets for image emotion regression tasks, demonstrating its effectiveness and superior performance. The emotion features extracted by the proposed multi-stage network exhibit strong discriminative power in the continuous dimensional space.

Emotion regression tasks have a more complex emotional space than image emotion classification tasks. Models must extract finer and more accurate emotion features for specific dimensional score predictions. The proposed method leverages a three-stage visual perception process to fully exploit the information in images. It simultaneously associates objective visual features with abstract emotional features through attribute perception tasks. The multi-stage method effectively addresses the challenges of feature learning in the complex continuous dimensional emotion space. Therefore, the proposed method enhances image emotion feature extraction through entity and attribution information, with learned features demonstrating strong discriminative power even in more challenging continuous emotion spaces.

### Image Emotion Distribution Learning

In this section, we compare the proposed method to related works on the task of image emotion distribution learning. The related works include problem transformation PT-SVM and PT-Bayes (Geng 2016), algorithm adaptations AA-kNN and AA-BP(Geng 2016), specialized algorithms SA-IIS and SA-BFGS (Geng 2016), deep convolutional networks CNNR (Peng et al. 2015), JCDL (Yang, She, and Sun 2017), WSCN (Yang et al. 2018), PDAN (Zhao et al. 2019), SSDL (Xiong et al. 2019), CSR (J. Yang, Li, et al. 2021), EAD (Xu and Wang 2021), and SAMN (Yang et al. 2022). CSR (J. Yang, Li, et al. 2021) establishes the distance relationship between emotion labels in the emotion space, SAMN (Yang et al. 2022) is an evaluation memory model based on deep convolutional networks, and EAD (Xu and Wang 2021) learns the correlations between emotion labels using graph convolutional networks. Distribution learning requires the model to capture correlations among different labels. Unlike these methods, the proposed method introduces emotion dependencies through fine-grained emotion distribution learning in the attribute perception process, enabling the model to learn the distribution of image emotions.

The experimental results, as shown in Tables [[tab:comparison3]](#tab:comparision3) and [[tab:comparison4]](#tab:comparision4), demonstrate that the proposed method achieves state-of-the-art performance on both the Flickr\_LDL and Twitter\_LDL data sets. As an example, the proposed method achieves KL divergences of 0.34 on both data sets, which are reduced by 0.06 and 0.09 compared to the next-best-performing method, SAMN. One intriguing observation is that AA-kNN consistently outperforms other algorithms on both data sets when evaluated using Clark and Canberra measures. This superiority can be attributed to AA-kNN’s ability to predict the distribution of emotions by considering the nearest k samples, enabling it to effectively capture emotions with low probabilities. Consequently, AA-kNN demonstrates enhanced performance on the Clark and Canberra metrics (Xiong et al. 2019). However, it is worth noting that AA-kNN exhibits poorer results on other evaluation metrics and does not attain a remarkable average ranking compared to the alternative algorithms. Overall, the proposed method achieves the best performance on most measures, indicating its superior performance in the task of emotion distribution learning.

Compared to the EAD method, which learns emotion correlations using graph convolutional networks, the proposed method utilizes attribute pseudo-labels as the correlation information between emotions, simplifying model learning and optimization. Compared to the CSR method, which is based on the distance relationship between emotion labels, the proposed method provides distributions of 66 emotion attribute words, enabling the model to learn more fine-grained information and generalize to distribution learning tasks with a larger number of label classes. Compared to the SAMN method, which is based on evaluation memory structures, the proposed method simplifies the evaluation process by perceiving attributes of entities in the image, eliminating the need for large-scale memory storage and resulting in a simpler model structure with fewer parameters.

In summary, the proposed method achieves good performance on the task of image emotion distribution learning and has the potential to perform well on fine-grained emotion label tasks.

## Visualization

Due to the more intuitive representation of image emotions conveyed by categorical labels compared to continuous dimensional labels and emotion distributions, we perform visualization analysis specifically for the task of image emotion classification. Figure [[fig:demo]](#fig:demo) illustrates several example outcomes of the multi-stage visual perception process. Each entity is accompanied by annotations denoting its category, attribute, and emotional contribution value. The figure demonstrates the commendable entity recognition performance of the proposed method, with accurate predictions for all entities displayed. Regarding attribute perception, the proposed method exhibits impressive analysis capabilities. In the "anger" and "fear" images, the model accurately predicts the emotional attributes for each individual. However, certain deviations in attribute prediction exist. For instance, in the "surprise" image, the attribute with the highest probability for the man is "sorrowful." Nevertheless, the influence of attribute misidentification is mitigated by the attribute distribution prediction. Furthermore, individual entities may struggle to fully express the overall emotion of an image. In the "surprise" image, for instance, the three entities exhibit weak emotional associations with surprise, highlighting the significance of global image semantics. The proposed method achieves a comprehensive understanding of global image semantics by analyzing scene attributes. Additionally, the attention mechanism of the model effectively emphasizes important information concerning each entity’s contribution to emotion prediction. For instance, the mournful teddy bear on the left side of the "sad" image holds a weight of 0.62, signifying its pivotal role in conveying the image’s emotional content. In summary, these examples substantiate the efficacy of the proposed method in visual content perception and the attention mechanism for emotional processing.

# Conclusion

To bridge the gap between visual content and emotional semantics, we propose a multi-stage perception network that draws inspiration from the emotion perception process. The network comprises entity, attribute, and emotion perception modules. Each module integrates features from the preceding module and deep visual features using an attention mechanism. To extract informative visual content, we employ an image segmentation model and the CLIP model to generate entity and emotion attribute labels, which are used to supervise the learning of the entity and attribute perception modules, respectively. As a result, the proposed multi-stage perception network progressively learns entities, attributes, and emotions in images, circumventing the direct learning of emotional semantics from visual signals. This approach effectively tackles the affective gap by breaking it down into manageable components.