Probing Sentiment-Oriented Pre-Training Inspired by Human Sentiment Perception Mechanism

Tinglei Feng* Jiaxuan Liu* Jufeng Yang†
TMCC, College of Computer Science, Nankai University, China
tinglyfeng@163.com, jxliu1999@163.com, yangjufeng@nankai.edu.cn

Abstract

*Pre-training of deep convolutional neural networks (DCNNs) plays a crucial role in the field of visual sentiment analysis (VSA). Most proposed methods employ the off-the-shelf backbones pre-trained on large-scale object classification datasets (*i.e.*, ImageNet). While it boosts performance for a big margin against initializing model states from random, we argue that DCNNs simply pre-trained on ImageNet may excessively focus on recognizing objects, but failed to provide high-level concepts in terms of sentiment. To address this long-term overlooked problem, we propose a sentiment-oriented pre-training method that is built upon human visual sentiment perception (VSP) mechanism. Specifically, we factorize the process of VSP into three steps, namely stimuli taking, holistic organizing, and high-level perceiving. From imitating each VSP step, a total of three models are separately pre-trained via our devised sentiment-aware tasks that contribute to excavating sentiment-discriminated representations. Moreover, along with our elaborated multi-model amalgamation strategy, the prior knowledge learned from each perception step can be effectively transferred into a single target model, yielding substantial performance gains. Finally, we verify the superiorities of our proposed method over extensive experiments, covering mainstream VSA tasks from single-label learning (SLL), multi-label learning (MLL), to label distribution learning (LDL). Experiment results demonstrate that our proposed method leads to unanimous improvements in these downstream tasks.* ***Our code is released on https://github.com/tinglyfeng/sentiment_pretraining***.

1. Introduction

Visual sentiment analysis aims to understand the sentiment embedded in an image, which gradually becomes a critical computer vision task that enables numerous applications from opinion mining [45], entertainment assistance [5], to business intelligence [18]. Given an image, the main goal of VSA is to recognize the emotion induced by viewers, providing either the categorical emotion states (CES) [9, 30] or dimensional emotion space (DES) [23, 41] representations. Traditional methods proposed for VSA normally involve extracting sentiment-related hand-crafted features like line directions [48], textures and colors [30], *etc*. These features are then sent to a classifier *e.g.*, a support vector machine (SVM) to predict the emotional states. However, due to affective gap [15], the low-level features can hardly meet the high-level attributes requirement of VSA, thus resulting in relatively unsatisfying performance.

Entering the deep learning era, DCNNs are now the dominant tools applied to various computer vision tasks, such as image classification, object detection, *etc*. Blessed with impressive high-level feature extraction capabilities, DCNNs have demonstrated superior advantages for modern VSA proven by a lot of milestone works [3, 50, 56]. Beneath the success, many may ignore one important factor that largely determines the performance of VSA, saying the pre-trained model. Due to the data-hungry nature of DCNNs, initializing model parameters from models trained on large-scale datasets has been a go-to technique for most tasks to improve their generalization abilities. When it comes to VSA, the lack of data has been exacerbated by the arduous annotation process (every image needs to be annotated by multiple people due to the subjectivity of emotion), resulting in its especially heavy reliance on pre-training. In our experiments on FI dataset [57], the ResNet50 [16] pre-trained on ImageNet [8] outperforms the one trained from scratch by 20 percent in terms of accuracy, revealing the undeniable crucial role the pre-trained model plays in VSA.

Today's deep models proposed for VSA are mostly initialized from models pre-trained on ImageNet to achieve satisfactory performance [59]. However, different from many other computer vision tasks that mainly depend on objective semantics, VSA requires a relatively higher level of understanding of an image. Therefore, pre-training only on ImageNet which is specially designed for object classification may not be the best practice for VSA.

In this paper, we argue that the models pre-trained on Ima-

* Equal contribution.
† Corresponding author.

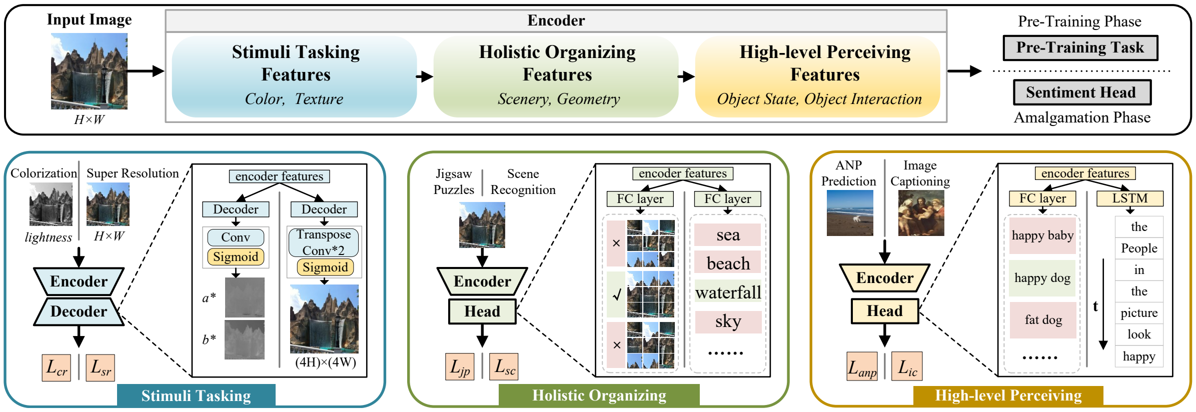

Figure 1. **Overview of our pre-training method.** We split a CNN backbone into three stages, each of which is responsible for extracting features corresponding to a certain VSP step. To fully excavate sentiment-related knowledge in terms of each step, a total of three models are separately trained to perform our elaborated tasks shown at the bottom.

geNet fail to achieve sentiment-related initial states to relieve the burden of learning sentiment representations from limited data. Also, due to the psychological and physiological nature of VSA, we believe that only if we fully understand how human sentiment is internally constructed can we thoroughly unveil the potential of VSA pre-training. Therefore, our proposed pre-training method is built upon human visual sentiment perception mechanism. Summarized from numerous existing research in the field of psychology and neuroscience [24, 26], we factorize the process of VSP into three steps in chronological order: 1) Stimuli Taking (ST): the procedure starts with the retina receiving light signals composed of colors and textures [29]. 2) Holistic Organizing (HO): the second step taking place in the primary visual cortex (V1) of our brain is to construct a whole map determining the overall context and global organization of scene [10, 43]. 3) High-level Perceiving (HP): the other parts of our brain help us separate the main objects from ambient light and build our high-level awareness [13, 19, 39].

Inspired by these theories, we build our pre-training framework by instructing the DCNNs to mimic the behavior of humans. In this work, we separately perform three groups of pre-training tasks, each of which is corresponding to one VSP step and is intentionally excavated the key sentiment features. To fully leverage the sentiment knowledge learned from pre-trained models, we then elaborate an amalgamation strategy to effectively distillate their abilities into a single target model. The amalgamation process is performed by squeezing the gap between the target model and sentiment-aware pre-trained models on both the logits and features at various levels. Moreover, the pre-trained models still participate in the whole downstream training, which further unleashes the potential learning abilities of DCNNs to accommodate the specialties of training data. We apply our method to multiple downstream VSA tasks including single-label learning, multi-label learning, and label distribution learning. Extensive experiments have demonstrated favorable improvements from our proposed pre-training method.

Our contributions are three-fold. 1) We propose a sentiment-oriented pre-training method to separately train a total of three models, each of which is dedicated to mimicking the human sentiment perception mechanism through performing pre-training tasks. 2) We devise an amalgamation strategy to aggregate the sentiment-discriminated knowledge from pre-trained models into a single target model during training downstream tasks, yielding favorable performance gains. 3) We conduct extensive experiments on various backbones and diverse VSA datasets. The experiment results demonstrate that our proposed method can unanimously improve the performance of a wide variety of VSA tasks.

2. Related Work

2.1. Visual Sentiment Analysis

VSA which aims to analyze the emotion induced by humans looking at an image attracts numerous researchers. Early pioneers study VSA by combining handcrafted features. In this milestone [28], 70 types of features revolving around colors and 23 types of features related to texture have shown profound impacts on visual sentiment. In addition, mid-level image attributes like scene and geometry have demonstrated their significance in image process [35], including the field of VSA proved by plenty of works [27, 38, 47]. Entering the deep learning era, today's methods based on deep neural networks have the capability to provide a significant high-level understanding of images, yielding numerous works [3, 52–54] that consistently set new state-of-the-art VSA performance records [20, 46]. In this work, we suggest that with the development of machine learning techniques, the evolution of VSA methods is similar to how visual signal is progressively perceived step by step from the retina to the cortex, thus we devise our pre-training method by mimicking the VSP mechanism to imitate human behavior.

2.2. Self-Supervised Learning

Having the advantages of training from artificially generated supervision signals, self-supervised learning has been attracting enormous attentions [11, 12, 14, 22, 31, 58]. Zhang *et al.* [58] first convert a photograph into a gray-scale image and then used it as input to network in order to predict a plausible colorful version image. Similar to this, Ledig *et al.* [22] propose to reconstruct a fine-grained up-sampled image from a low-resolution one, yielding a pre-trained model that is comparable to the supervised method on some downstream tasks. Considering that images contain rich context and layout information, *e.g.*, relative positions among different objects, [31] proposes to perform self-supervised learning from solving Jigsaw puzzles. They first crop an image into multi-patches and send them to networks to predict their permutation types among all preset possibilities. In this work, we employ the aforementioned self-supervised learning methods as part of our pre-training tasks.

2.3. Knowledge Distillation and Amalgamation

Knowledge distillation (KD) [17] is a widely researched technique mainly used to improve the performance of a lightweight model by learning from a fancier and larger teacher model. On the other hand, Knowledge amalgamation (KA) [55] aims to aggregate the prior multi-domain or multi-modality knowledge from several models to a single target model. While KA is similar to KD, it holds its own unique specialties and applications. First, the apparent distinction is that KD normally involves only two models (*i.e.*, student model and teacher model) [33] while amalgamation call for the participation of multiple models. Second, KD is mostly applied to two models performing the same task [51, 61], but KA commonly works in scenarios intertwined with multi-task learning [42]. Given such differences in a macro perspective, the implementation details beneath the two paradigms are similar. The learning procedure from one model to another model is mainly performed by squeezing the distance between either their intermediate feature representations or the logits from the last layer. In our work, we implement our amalgamation strategy through features-level and logits-level regularization.

3. Method

3.1. Overview

As demonstrated by Fig. 1, our pre-trained method is built upon a multi-task framework. The components of each model can be divided into two categories: backbone blocks (encoder) and heads (decoder). The former basically consists of multiple consecutive CNN blocks for feature extraction. Fed with the captured features, each head in our proposed method is corresponding to executing a specific task. The labels used to supervise these tasks are either from the original dataset or artificially generated, which means we employ two common learning paradigms, namely fully-supervised learning (FSL) and self-supervised learning (SSL). In the pre-training phase, three models have been separately trained on the tasks that are intended to imitate different human VSP steps. Fig. 2 demonstrates how we amalgamate the prior knowledge learned from pre-training into a target model during performing downstream tasks. We will detail our pre-training and amalgamation method below.

3.2. Pre-Training for Stimuli Taking

The visual stimuli occur in the first step of VSP when the light signal passes through the retina of the eyes, where the signal can be decomposed into two basic image attributes of colors and textures. The two attributes have also been proved to have direct impacts on deciding perceived sentiment [60]. A rule of thumb is that images accompanying bright color and harmonious texture usually bring contentment while those with murky backgrounds and irregular lines tend to make most of us melancholy. As shown in Fig. 3, the pleasant weather depicted by the bright color and soft sand naturally brings the observer a positive atmosphere. While assuming that on rainy days the dark sky and muddy road inevitably make us depressed. In this work, we propose to learn these two kinds of features via self-supervised learning.

• **Colorization** Given a gray-scale image as input, the goal of image colorization is to predict the plausible color version of that image. We follow this milestone work [58] to construct the colorization task. For a *RGB* image $\boldsymbol{I} \in R^{H\times W\times 3}$, we first convert it to *LAB* color space, where $\boldsymbol{L}$ correlates with lightness, $\boldsymbol{A}$ and $\boldsymbol{B}$ reflect colors. We then take the $\boldsymbol{L}$ space matrix $\boldsymbol{X}^{cr} \in R^{H\times W\times 1}$ as input to the backbone network, and the *AB* spaces matrix $\boldsymbol{Y}^{cr} \in R^{H\times W\times 2}$ as the target of our prediction. In our practice, directly regressing two dense pixel maps (*i.e.*, $\boldsymbol{Y}^{cr}$) yields favorable results in terms of pre-training. Thus we formulate the loss of colorization as $\boldsymbol{L}_{cr}$:

$$\boldsymbol{L}_{cr} = \frac{1}{HWC}\sum_{i,j,k}(\boldsymbol{Y}^{cr}_{i,j,k} - \hat{\boldsymbol{Y}}^{cr}_{i,j,k})^2, \tag{1}$$

where $\hat{\boldsymbol{Y}}^{cr}$ is the predictions from colorization head.

• **Super Resolution** Image super resolution (ISR) aims to recover the finer texture details from up-scaling an image at low resolution [22]. We integrate this task into our pre-training with a similar architecture to Tong *et al.* [44]. It can be regarded as a fully convolutional network (FCN) where the backbone is used to encode image futures and the ISR head is the decoder to restore finer details. We employ deconvolutional layers to simultaneously upscale and map the low-resolution features to a high-resolution

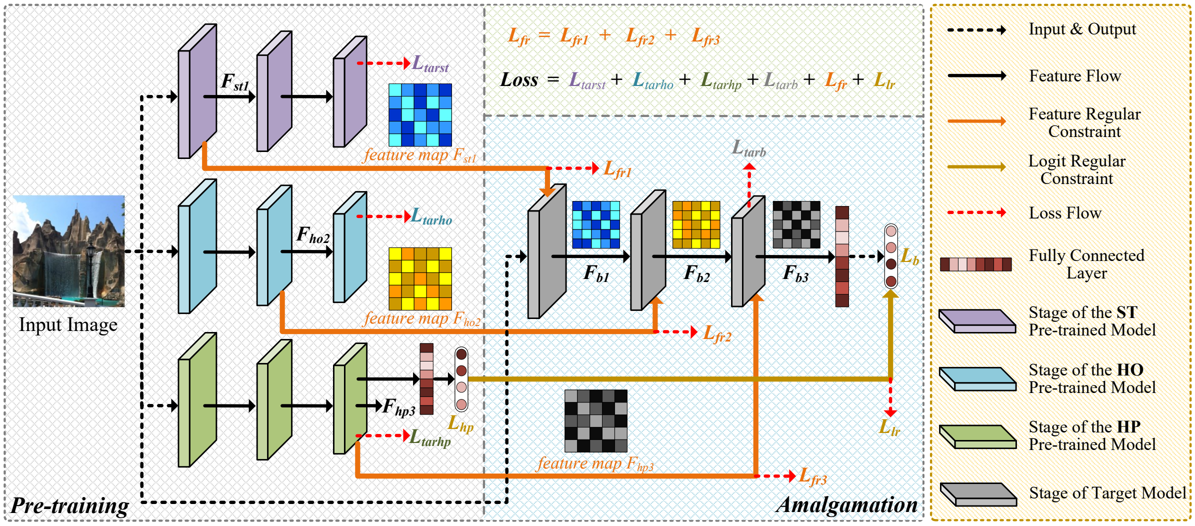

Figure 2. **The pipeline of our proposed amalgamation strategy.** There are four backbones participating in training at the same time, *i.e.*, the three pre-trained models and a target model. The knowledge transfer is implemented by regularization on intermediate feature maps and output logits. The feature constraint is applied between the three pre-trained models and the target model while the logits constraint is only adopted between the HP pre-trained model and the target model.

restored image. To train the backbone network and ISR head, we adopt MSE loss computed between the original image $\boldsymbol{X}^{sr} \in R^{H\times W\times C}$ and the predicted image $\hat{\boldsymbol{Y}}^{sr} \in R^{H\times W\times C}$, denoted as:

$$\boldsymbol{L}_{sr} = \frac{1}{HWC}\sum_{i,j,k}(\boldsymbol{X}^{sr}_{i,j,k} - \hat{\boldsymbol{Y}}^{sr}_{i,j,k})^2. \qquad (2)$$

3.3. Pre-Training for Holistic Organizing

The second step of VSP taking place in the primary visual cortex is responsible for building an overview of the whole reception field and constructing a holistic map [10,43]. Grounded into certain image attributes, we suggest that the geometry and scenery reflecting the global organization and holistic context of an image are the best choices to be the pre-training pretext. As shown in Fig. 3, the vast sea and rule of thirds have a huge impact on determining our feelings when we have a glance at this picture. To improve the abilities of DCNNs to extract these two kinds of features, we employ both the FSL and SSL paradigms.

- **Scene Recognition** Since our pre-training method is based on Places365 dataset [62], a large-scale scene recognition dataset spanning more than three hundred categories, it is intuitive that we can directly supervise our model with the given labels. Here we can simply treat the scene recognition task as a classification problem. Given an input image $\boldsymbol{X}^{sc} \in R^{H\times W\times 3}$ and its ground truth $\boldsymbol{Y}^{sc} \in \{0,1\}^C$, where C is the total number of scene categories, the loss for scene recognition is denoted as:

$$\boldsymbol{L}_{sc} = -\sum_{i=1}^{C}\mathbf{Y}^{sc}_i log(\hat{\mathbf{Y}}^{sc}_i), \qquad (3)$$

where $\hat{\boldsymbol{Y}}^{sc}$ is the possibility for each class output from the softmax layer. All the following three tasks can be formulated as classification problems.

- **Jigsaw Puzzles** Solving Jigsaw Puzzles is proven to be an effective self-supervised task in computer vision [31]. Given an image, we first uniformly crop it to multiple small patches (*e.g.*,3×3) and then randomly shuffle and reconstruct them to a new disordered image. The goal of the network is to infer the correct position of each patch by learning the relative structure and geometry relationships within intra- or inter-objects. We can formulate the Jigsaw puzzles as a classification task where we can directly predict the correct permutation among limited possibilities. However, given 3×3 patches from an image, there exists $362880 = 9!$ combinations, which are too large to be accurately recognized by deep models. Following [31], we only pick limited permutations that have relatively large hamming distances.

3.4. Pre-Training for High-level Perceiving

Human sentiment is intrinsically a high-level concept. In the third step of VSP, there are specialized partitions of the human brain that deal with higher-level information, *e.g.*, emotions in verbal logic [19], and the salience area [39]. As demonstrated in Fig. 3, the decisive factor of why we humans feel pleasure when looking at this image is that we can easily tell who is the protagonist (the little girl), how is she (happily smiling) and infer from the whole context that what she is doing (playing with sand). In this paper, we propose to incorporate two essential high-level perceiving modalities in the field of VSP, saying adjective-noun pairs (ANP) and image caption(IC).

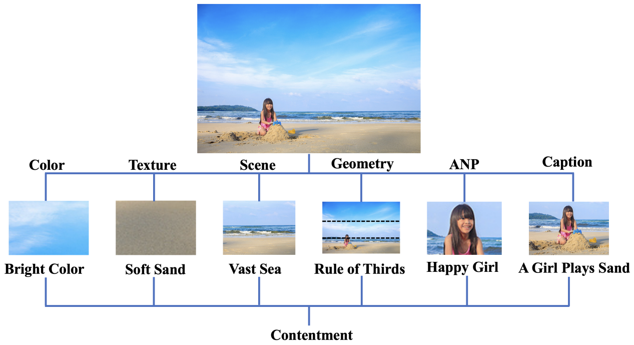

Figure 3. Demonstration of how an image is perceived based on image attributes from each VSP step. In this example, color and texture from ST, scene and geometry from HO, ANP and caption from HP contribute together to bringing observer contentment.

- **ANP Prediction** ANP stands for adjective-noun pairs. In this paper, we employ the commonly used VSO [2] dataset. The dataset comprises 1200 ANPs, which is built according to the psychological theory of Plutchik's Wheel of Emotions [37]. Each ANP is ensured to reflect a strong sentiment and link to emotions, for instance, beautiful flowers, disgusting food, angry cat, *etc*. The purpose of adopting ANP prediction is two-fold. First, the nouns here are equivalent to objects, which is a high-level visual concept that can be essential for VSA. For example, the "flower" normally gives viewers positive emotions while the "shark" tends to make us fear. One may argue that ImageNet has already provided sufficient object semantics, but the pitfall here is that sometimes the same object could induce different feelings. A piece of simple evidence is that "vicious dog" can intimidate viewers while "happy dog" brings us contentment. This phenomenon introduces the second purpose of adopting ANP, which is the significance of adjectives in ANP that describe the states of objects. In the VSO dataset, the same object with a different adjective (*i.e.*, emotion-related state) is divided into different categories. We argue that such taxonomy can encourage the DCNNs to learn features that help to distinguish different states, which is more suitable for VSA than simply employing the ImageNet dataset.

- **Image Captioning** The task of image captioning is to generate a description for an image based on its content. Image Captioning is usually considered a high-level task where the model should not only determine which objects are in an image but also reveal the relationships between different objects and express them with natural language. While in the field of VSA, learning to understand how an object or person interacts with others often plays an important or even crucial role in recognizing emotions. For instance, there are both two images depicting a wife and a husband. The content in the first one is "The wife is hugging her husband" while another one shows "The husband is abusing his wife". Apparently, the former makes viewers pleasant but the second rises our anger. In this work, to excavate the relationships between objects, we employ the Image Captioning task as one of our pre-training tasks. Moreover, rather than choose the normally used captioning dataset like COCO Captions [6], we prefer the more emotion-related ArtEmis [1]. Each image in this dataset is provided with multiple emotion tags like awe, fear, excitement, *etc*, along with each of which is a description from annotators for expressing why they tend to have these feelings when watching it. For instance, *contentment* is given to an artwork portraying a bunch of people due to "These people seem to be getting along and happy to be with one another, which makes me feel calm and accepted". We add the Image Captioning task to our network by adding an LSTM head to the end of the backbone.

3.5. Sentiment Amalgamation

Once the backbone is pre-trained on the aforementioned tasks, a crucial question left is how we utilize the sentiment-oriented model parameters. A simple solution comes that we first train all the tasks simultaneously on the same backbone, and then directly load the produced pre-trained model while training downstream tasks. However, this strategy does not bring us satisfactory performance gains in our experiments. According to [40], the paradigm of multi-task learning generally works based on the assumption that the performed tasks are tightly related to each other. In other words, MLL can hardly be effective when some tasks are irrelevant or even conflict with others. In our scenario, the six tasks are responsible for learning prior sentiment knowledge from the perspective of VSP mechanism. Since every two tasks are designed for one specific VSP step, it is not applicable to simply train all the tasks at once since the learning direction for each task is not the same and may contradict others (*e.g.*, the low-level colorization and high-level captioning).

To solve this problem, we propose a multi-model amalgamation strategy. As shown in Fig. 2, we distribute the six tasks to three separated models, with model $\mathcal{M}_{st}$ for colorization and super resolution, model $\mathcal{M}_{ho}$ for scene recognition and jigsaw puzzles, model $\mathcal{M}_{hp}$ for ANP prediction and image captioning. In this way, tasks in the same model are highly related and similar, so that they can freely learn specific features in terms of their VSP step to benefit each other, without the disturbance from other non-related tasks. Further, we propose a knowledge amalgamation method to effectively transfer various levels of prior knowledge into a single target model. As demonstrated in Fig. 2, a total of 4 models participate in the training of downstream tasks. One of these models is the base model $\mathcal{M}_b$ and the other three are pre-trained models. We separate each backbone into three stages and denote the intermediate features from i_{th} stage in the base model as $\boldsymbol{F_{bi}} \in \mathbb{R}^{\mathrm{C_i \times H_i \times W_i}}$, and $\boldsymbol{F_{sti}}$, $\boldsymbol{F_{hoi}}$, $\boldsymbol{F_{hpi}} \in \mathbb{R}^{\mathrm{C_i \times H_i \times W_i}}$ for pre-trained models, where C_i, H_i, W_i represent the channel, width, and height of feature maps outputted from i_{th} stage. Similarly, the logits output from each model is denoted as $\boldsymbol{L_b}$, $\boldsymbol{L_{st}}$, $\boldsymbol{L_{ho}}$, $\boldsymbol{L_{hp}} \in \mathbb{R}^{\mathrm{N}}$, where N is the number of sentiment classes.

We observe that the feature extraction by forwarding images from shallow layers to deep layers in DCNNs exactly resembles the pipeline of VSP: In the earlier period, the entrance layers of DCNNs (stage 1) aim to extract features from the basic colors and textures. Based on these local features, the following layers (stage 2) with expanded reception fields can progressively capture the global context. Finally, the highest layers (stage 3) and classifiers are responsible for high-level semantic and sentiment understanding. In conclusion, the VSP process from the eyes to the brain is similar to the feature flow in DCNNs layers. Therefore, to effectively assimilate affluent prior sentiment knowledge in terms of each VSP step, our amalgamation process is performed between the pre-trained models and the corresponding stage of the target model, where the knowledge transfer is achieved by two types of regularization. The first is the feature regularization performed on the output features from every intermediate stage in a backbone model. Specifically, we assign three sets of this regularization on the first stage between $\mathcal{M}_b$ and $\mathcal{M}_{st}$, the second stage between $\mathcal{M}_b$ and $\mathcal{M}_{ho}$, and the third stage between $\mathcal{M}_b$ and $\mathcal{M}_{hp}$:

$$\mathcal{L}_{fr} = \frac{1}{C_1 H_1 W_1}||\boldsymbol{F_{b1}} - \boldsymbol{F_{st1}}||_F^2 + \\ \frac{1}{C_2 H_2 W_2}||\boldsymbol{F_{b2}} - \boldsymbol{F_{ho2}}||_F^2 + \frac{1}{C_3 H_3 W_3}||\boldsymbol{F_{b3}} - \boldsymbol{F_{hp3}}||_F^2, \tag{4}$$

where $||\boldsymbol{F}||_F = \sqrt{\sum_{i=i}^{C} \sum_{j=1}^{H} \sum_{k=1}^{W} F_{ijk}^2}$ is the Frobenius Norm of a matrix.

Another regularization is performed on the logits outputted from the last fully connected layer. The difference here is that we only add constraint between the $\boldsymbol{L_b}$ from $\mathcal{M}_b$ and $\boldsymbol{L_{hp}}$ from $\mathcal{M}_{hp}$, which can be formulated as:

$$\mathcal{L}_{lr} = ||\boldsymbol{L_b} - \boldsymbol{L_{hp}}||_2^2. \tag{5}$$

This special setting is derived from that the logits used to make the final decision for better predictions are naturally prone to higher and more abstract features, and we will show detailed experimental analysis in the following section. By adding the two above constraints on both the features and logits, we can transfer the base model with comprehensive extra sentiment knowledge that it can hardly learn from the limited downstream data.

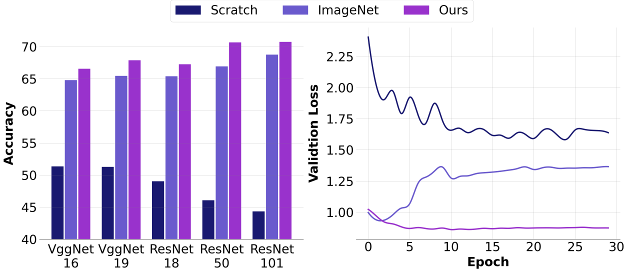

Figure 4. Illustration of the effectiveness of different pre-training strategies. The left figure presents the accuracy produced from various backbones on the FI dataset. The curve on the right is loss movement of ResNet50 with the increase of training epochs.

Finally, an indispensable part of training is the target loss between the predictions and ground truth labels for each specific downstream task. Instead of assigning the target loss $\mathcal{L}_{tarb}$ for only the base model, we additionally back-propagate the prediction error $\mathcal{L}_{tarst}$, $\mathcal{L}_{tarho}$, $\mathcal{L}_{tarhp}$ of all the pre-trained models. In this way, we can adaptively adjust the pre-trained parameters to accommodate different datasets in distinct domains. The total loss of the whole amalgamation is defined as:

$$L = \mathcal{L}_{fr} + \mathcal{L}_{lr} + \mathcal{L}_{tarb} + \mathcal{L}_{tarst} + \mathcal{L}_{tarho} + \mathcal{L}_{tarhp}. \tag{6}$$

In our empirical experiments, the results are robust to different sets of weights assigned to each loss. Thus we simply omit the balance weights.

4. Experiments

4.1. Datasets and Model Settings

• **Datasets** All the tasks in ST and HO pre-training are performed on Places365 [62] dataset, while the ANP Prediction and Image Captioning in HP pre-training are conducted on VSO [2] and ArtEmis [1] datasets respectively. To verify the superiority of our proposed method, we conduct experiments across three learning tasks (*i.e.*, SLL, MLL, LDL) over 7 VSA datasets. More specifically, We perform our SLL experiments on FI [57], UnBiasedEmo [32] datasets, MLL experiments on Emotic [21], Emotion6 [36] datasets, LDL experiments on Emotion6 [36], Abstract [28] datasets.

• **Implementation Details** Our experiments are based on Pytorch [34] framework running on two NVIDIA RTX 3090 GPUs. For the pre-training phase, we only employ the basic data augmentation strategy where we first resize a given image to 256 × 256 and then randomly crop it to 224 × 224, finally a horizontal flip with probability 0.5 is applied to this image. During training, we set the batch size to 2^n and optimize our model with stochastic gradient descent (SGD), where n is the maximum number to train models without exceeding the limit of GPU memory. The initial learning rate is set to 0.01. For HO and HP pre-training, we set the total epochs to 30 and divide the learning rate by 10 every 10 epoch. While the total epochs and decay interval in the ST pre-training are 12 and 4. During the training of downstream tasks, we adopt the same simple data augmentation strategy. We set the initial learning rate, total epochs, and decay interval to 0.001, 30, and 10, respectively.

Table 1. Results of several classic backbone networks on single-label learning datasets including FI, UnBiasedEmo (UB), multi-label learning datasets including Emotic (EM), Emotion6 (E6), and label distribution learning datasets including Emotion6 (E6), Abstract (AB). The numbers on the left are from models initialized from ImageNet while the numbers on the right are from our proposed method.

	Dataset	Metric	Backbone				
			Vgg16	Vgg19	ResNet18	ResNet50	ResNet101
SLL	FI [57]	Acc ↑	0.648 → 0.666	0.655 → 0.679	0.654 → 0.673	0.670 → 0.707	0.688 → 0.708
	UB [32]	Acc ↑	0.773 → 0.839	0.770 → 0.846	0.816 → 0.836	0.821 → 0.839	0.823 → 0.859
MLL	EM [21]	Hamming ↓	0.151 → 0.147	0.149 → 0.145	0.163 → 0.145	0.155 → 0.154	0.136 → 0.137
		Ranking ↓	0.159 → 0.155	0.158 → 0.154	0.161 → 0.154	0.145 → 0.143	0.155 → 0.152
		MicroF1 ↑	0.155 → 0.199	0.156 → 0.219	0.153 → 0.214	0.218 → 0.219	0.204 → 0.226
		MacroF1 ↑	0.024 → 0.038	0.026 → 0.042	0.026 → 0.040	0.041 → 0.043	0.046 → 0.049
	E6 [36]	Hamming ↓	0.260 → 0.180	0.253 → 0.171	0.248 → 0.187	0.256 → 0.221	0.158 → 0.157
		Ranking ↓	0.294 → 0.226	0.295 → 0.229	0.282 → 0.236	0.191 → 0.160	0.216 → 0.216
		MicroF1 ↑	0.809 → 0.846	0.808 → 0.845	0.817 → 0.842	0.831 → 0.849	0.848 → 0.849
		MacroF1 ↑	0.726 → 0.825	0.718 → 0.823	0.776 → 0.817	0.798 → 0.827	0.825 → 0.827
LDL	E6 [36]	Chebyshev ↓	0.335 → 0.322	0.337 → 0.326	0.278 → 0.276	0.252 → 0.251	0.259 → 0.253
		Clark ↓	1.669 → 1.660	1.672 → 1.662	1.636 → 1.629	1.614 → 1.615	1.621 → 1.618
		Canberra ↓	3.768 → 3.721	3.775 → 3.732	3.620 → 3.587	3.516 → 3.514	3.533 → 3.533
		KL ↓	0.637 → 0.597	0.642 → 0.608	0.466 → 0.464	0.403 → 0.402	0.405 → 0.391
		Cosine ↑	0.697 → 0.716	0.694 → 0.710	0.786 → 0.788	0.822 → 0.823	0.818 → 0.835
		Intersection ↑	0.553 → 0.573	0.551 → 0.568	0.634 → 0.636	0.666 → 0.669	0.684 → 0.687
	AB [28]	Chebyshev ↓	0.268 → 0.266	0.267 → 0.267	0.279 → 0.259	0.256 → 0.247	0.258 → 0.249
		Clark ↓	1.663 → 1.658	1.660 → 1.660	1.714 → 1.677	1.662 → 1.644	1.653 → 1.655
		Canberra ↓	3.950 → 3.930	3.928 → 3.924	4.086 → 3.940	3.936 → 3.835	3.886 → 3.880
		KL ↓	0.580 → 0.568	0.573 → 0.571	0.702 → 0.570	0.568 → 0.517	0.554 → 0.536
		Cosine ↑	0.702 → 0.710	0.707 → 0.708	0.642 → 0.711	0.707 → 0.743	0.719 → 0.731
		Intersection ↑	0.579 → 0.585	0.584 → 0.585	0.553 → 0.595	0.587 → 0.614	0.597 → 0.605

Table 2. Comparison with SSL methods on FI dataset.

Method	SimCLR [4]	SimSiam [7]	MoCoV3 [49]	A2MIM [25]	**Ours**
Accuracy	0.627	0.636	0.626	0.484	**0.707**

4.2. Effectiveness of Our Pre-Training Method

The foundation of our work is that pre-training is a crucial impetus for improving the performance of DCNNs on VSA. To first prove this assumption, we conduct experiments with a set of naive DCNNs initializing from scratch and another set initializing from ImageNet pre-trained models. The results are shown in Fig. 4. We are not surprised to see that the pre-trained models outperform the naive by a big margin, reaching approximately 20 percent improvements. An interesting observation is that the models with more elegant architecture (ResNet vs. VggNet) and higher capacity (ResNet101 vs. ResNet18) perform worse when training from scratch. Such a phenomenon is mainly attributed to the notorious overfitting problem due to a lack of data. That is why DCNNs are usually criticized for their huge requirement of tons of data. And their performances further deteriorate when it comes to data deficiency in the field of VSA. As the knowledge transferred from tasks such as object classification can save the model from exhausting learning general features, the problem is alleviated. However, the model performance can be further boosted if we provide models with more prior knowledge of sentiment. As shown in Fig. 4, our proposed pre-training method has the ability to transfer multi-level prior sentiment knowledge from the pre-trained model into the target model, thus it achieves consistent improvements over the ImageNet pre-training strategy.

From observing the loss curve on the right of Fig. 4, we can dive deeper to inspect why our method performs better. First, due to nearly no awareness of extracting image features, the loss of the model trained from scratch converges very slowly and heavily fluctuates during the whole training, remaining relatively higher loss at the end. Second, the model pre-trained on ImageNet presents impressive learning ability at the start of training, but sinks into severe overfitting because of lack of sentiment knowledge. Finally, the target model in our method can constantly learn sentiment-related features from not only the current dataset but also other mature models carrying affluent sentiment knowledge, hence it converges much faster and continues learning smoothly in the whole training process.

Compared to ImageNet pre-training method in downstream training, the total parameters involved in our method are quadrupled, and the CUDA time for training a single

Table 3. Ablation study of feature and logits regularization from different pre-trained models with ResNet50 on FI dataset.

Sentiment Stages			Feature Reg	Logits Reg
ST	HO	HP	Acc ↑	Acc ↑
–	–	–	66.95	66.95
✓	–	–	67.83	66.38
–	✓	–	67.89	69.50
–	–	✓	68.15	**70.20**
✓	✓	✓	**68.92**	68.83

Table 4. Ablation study of the effectiveness of target loss (TL) in our method on FI dataset.

Models	Vgg16	Vgg19	ResNet18	ResNet50	ResNet101
w/o TL	62.46	64.19	65.89	69.27	70.21
w/ TL	66.57	67.89	67.27	70.68	70.77

sample in our method (ResNet50) is about 39 ms while the ImageNet pre-training takes 6 ms. However, in terms of inference latency, a more important metric that evaluates computation consumption, our pure training-time techniques achieve better predictions without any expense of speed.

4.3. Unanimous Improvements over VSA Tasks

We conduct extensive experiments to verify that our pre-training method has a strong generalization ability to improve the performance of VSA on a total of three dimensions, *i.e.*, tasks, models, and datasets, as shown in Tab. 1. From the perspective of the task dimension, we can conclude that our pre-training strategy can simply increase the accuracy of single-label classification by 2 to 3 percent. When it comes to more fine-grained tasks like MLL and LDL, the results on multiple datasets prove that prior knowledge relating to sentiment can be also beneficial for recognizing more diverse and implicit emotions brought to humans. Focusing on the model dimension, *i.e.*, each column of Tab. 1, an obvious observation is that the more sophisticated and bigger models generally perform better on VSA like on other tasks. For all of these different backbones, we can reach a conclusion that our proposed method achieves unanimous improvements, which shows the robustness of our pre-training strategy regardless of model architecture or model size.

Further, we compare the proposed method with recent SSL methods [7, 25, 49, 49], where the training data is the same as ours. As shown in Tab. 2, our method outperforms others by a large margin, which attributes to our utilization of the characteristic of sentiment (*e.g.*, hierarchy) rather than the diversity of data distributions.

4.4. Ablation Study

To explore the effectiveness of each part in our method, we conduct extensive experiments with ResNet50 on the FI dataset. As shown in Tab. 3, we first investigate the impact of each set of feature amalgamation. The bare backbone without any prior knowledge related to sentiment can only reach an accuracy of 66.95%. Once we add the ST or HO feature regularization, the accuracy is boosted by approximately one percent. Moreover, we can observe that high-level perceiving regularization produces slightly more performance gains, which tells us that visual sentiment recognition relies more on the more advanced and abstract features. Nevertheless, these three kinds of sentiment knowledge from different perspectives are all crucial factors of VSA. And only if we fully utilize all the knowledge learned from each VSP step can we achieve the best results, which is proved by the last column. Different from the feature level, the ablation studies of logits regularization gives us another perspective of VSA. First, forcing the logits output of the base model to be similar to logits from only the ST pre-trained model can degenerate model performance. Second, employing all three regularization does not work as better as only performing the HP regularization. These facts imply that the logits used to make the final decision tend to correlate with the highest level of stimulation, which makes sense according to the VSP mechanism—that is, the first ST and second HO are mainly used for pre-processing low- and mid- level signals, as for high-level understanding, it is our brain's responsibility and functionality to perform sentiment perceiving and decision making. Finally, we also provide comparisons between models with and without target loss. As explained in Tab. 4, adding target loss of pre-trained models is an essential part of our amalgamation strategy. With target loss, the pre-trained model can flexibly adjust its model parameter to accommodate specifies of different datasets, thus narrowing the gap from the domain discrepancy between pre-training datasets and downstream datasets, and therefore, easing the training of the whole amalgamation process.

5. Conclusion

In this work, we explore the crucial yet long-term overlooked pre-training paradigm in the field of VSA. We first propose a sentiment-oriented pre-training strategy via both fully-supervised and self-supervised learning to excavate sentiment-specific features from the perspective of VSP mechanism. Next, an elaborated sentiment amalgamation strategy is proposed to effectively transfer all the prior sentiment knowledge into a single model. Finally, we verify the effectiveness of our method with various VSA tasks and diverse datasets. The limitation is that our amalgamation process is more computation-intensive compared to directly loading a pre-trained model. We will devote ourselves to solving this problem in the future.

6. Acknowledgments

This work was supported by the National Key Research and Development Program of China Grant (NO. 2018AA A0100400), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), and Fundamental Research Funds for the Central Universities.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021. 5, 6

[2] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 459–460, 2013. 5, 6

[3] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014. 1, 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 7

[5] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 367–376, 2014. 1

[6] 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. 5

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 7, 8

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 1

[9] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992. 1

[10] Edward F Ester, John T Serences, and Edward Awh. Spatially global representations in human primary visual cortex during working memory maintenance. *Journal of Neuroscience*, 29(48):15258–15265, 2009. 2, 4

[11] Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, et al. Task agnostic and task specific self-supervised learning from speech with lebenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[12] Lluis Gomez, Yash Patel, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4230–4239, 2017. 3

[13] Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27:649–677, 2004. 2

[14] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, HONG Lanqing, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, et al. Soda10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[15] Alan Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[17] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[18] Morris B Holbrook and John O'Shaughnessy. The role of emotion in advertising. *Psychology & Marketing*, 1(2):45–64, 1984. 1

[19] Khodijah Hulliyah, Normi Sham Awang Abu Bakar, and Amelia Ritahani Ismail. Emotion recognition and brain mapping for sentiment analysis: A review. In *2017 Second International Conference on Informatics and Computing (ICIC)*, pages 1–5. IEEE, 2017. 2, 4

[20] Guoli Jia and Jufeng Yang. S 2-ver: Semi-supervised visual emotion recognition. In *European conference on computer vision*, pages 493–509, 2022. 2

[21] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1667–1675, 2017. 6, 7

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photorealistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 3

[23] Joonwhoan Lee and EunJong Park. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13(5):1031–1039, 2011. 1

[24] Giada Lettieri, Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Paolo Papale, Monica Betta, Pietro Pietrini, and Luca Cecchetti. Emotionotopy in the human right temporoparietal cortex. *Nature communications*, 10(1):1–13, 2019. 2

[25] Siyuan Li, Di Wu, Fang Wu, Zelin Zang, Baigui Sun, Hao Li, Xuansong Xie, Stan Li, et al. Architecture-agnostic masked image modeling–from vit back to cnn. *arXiv preprint arXiv:2205.13943*, 2022. 7, 8

[26] Kristen A Lindquist, Tor D Wager, Hedy Kober, Eliza Bliss-Moreau, and Lisa Feldman Barrett. The brain basis of emotion: a meta-analytic review. *The Behavioral and Brain Sciences*, 35(3):121, 2012. 2

[27] Xin Lu, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. An investigation into three visual characteristics of complex scenes that evoke human emotion. In *2017*

Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 440–447. IEEE, 2017. 2

[28] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 83–92, 2010. 2, 6, 7

[29] Richard H Masland. The neuronal organization of the retina. *Neuron*, 76(2):266–280, 2012. 2

[30] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005. 1

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 69–84. Springer, 2016. 3, 4

[32] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 579–595, 2018. 6, 7

[33] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 3

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019. 6

[35] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012. 2

[36] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2015. 6, 7

[37] Robert Plutchik. A psychoevolutionary theory of emotions, 1982. 5

[38] Tianrong Rao, Min Xu, Huiying Liu, Jinqiao Wang, and Ian Burnett. Multi-scale blocks based image emotion classification using multiple instance learning. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 634–638. IEEE, 2016. 2

[39] Edmund T Rolls. Vision, emotion and memory: from neurophysiology to computation. In *International Congress Series*, volume 1250, pages 547–573. Elsevier, 2003. 2, 4

[40] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 5

[41] Harold Schlosberg. Three dimensions of emotion. *Psychological Review*, 61(2):81, 1954. 1

[42] Chengchao Shen, Mengqi Xue, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3504–3513, 2019. 3

[43] Frank Tong. Primary visual cortex and visual awareness. *Nature Reviews Neuroscience*, 4(3):219–229, 2003. 2, 4

[44] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international Conference on Computer Vision*, pages 4799–4807, 2017. 3

[45] Quoc-Tuan Truong and Hady W Lauw. Visual sentiment analysis for review images with item-oriented and user-oriented cnn. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1274–1282, 2017. 1

[46] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 218–227, 2022. 2

[47] Xiaohui Wang, Jia Jia, Jiaming Yin, and Lianhong Cai. Interpretable aesthetic features for affective image classification. In *2013 IEEE International Conference on Image Processing*, pages 3230–3234. IEEE, 2013. 2

[48] Wang Wei-ning, Yu Ying-lin, and Zhang Jian-chao. Image emotional classification: static vs. dynamic. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 7, pages 6407–6411. IEEE, 2004. 1

[49] Chen Xinlei, Xie Saining, and He Kaiming. An empirical study of training self-supervised visual transformers. *arXiv preprint arXiv:2104.02057*, 8, 2021. 7, 8

[50] C Xu, S Cetintas, KC Lee, and LJ Li. Visual sentiment prediction with deep convolutional neural networks (2014). *arXiv preprint arXiv:1411.5731*. 1

[51] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 3

[52] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021. 2

[53] Jufeng Yang, Dongyu She, Ming Sun, Ming-Ming Cheng, Paul L Rosin, and Liang Wang. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9):2513–2525, 2018. 2

[54] Xingxu Yao, Dongyu She, Sicheng Zhao, Jie Liang, Yu-Kun Lai, and Jufeng Yang. Attention-aware polarity sensitive embedding for affective image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1140–1150, 2019. 2

[55] Jingwen Ye, Xinchao Wang, Yixin Ji, Kairi Ou, and Mingli Song. Amalgamating filtered knowledge: Learning task-customized student from multi-task teachers. *arXiv preprint arXiv:1905.11569*, 2019. 3

[56] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Twenty-ninth AAAI Conference on Artificial Intelligence*, 2015. 1

[57] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 1, 6, 7

[58] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. 3

[59] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73, 2021. 1

[60] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[61] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Mingming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 4, 6