mathematics

MDPI

Article

Multi-Output Learning Based on Multimodal GCN and Co-Attention for Image Aesthetics and Emotion Analysis

Haotian Miao *, Yifei Zhang, Daling Wang and Shi Feng

School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China; zhangyifei@cse.neu.edu.cn (Y.Z.); wangdaling@cse.neu.edu.cn (D.W.); fengshi@cse.neu.edu.cn (S.F.)
* Correspondence: huhumht@gmail.com

Abstract: With the development of social networks and intelligent terminals, it is becoming more convenient to share and acquire images. The massive growth of the number of social images makes people have higher demands for automatic image processing, especially in the aesthetic and emotional perspective. Both aesthetics assessment and emotion recognition require a higher ability for the computer to simulate high-level visual perception understanding, which belongs to the field of image processing and pattern recognition. However, existing methods often ignore the prior knowledge of images and intrinsic relationships between aesthetic and emotional perspectives. Recently, machine learning and deep learning have become powerful methods for researchers to solve mathematical problems in computing, such as image processing and pattern recognition. Both images and abstract concepts can be converted into numerical matrices and then establish the mapping relations using mathematics on computers. In this work, we propose an end-to-end multi-output deep learning model based on multimodal Graph Convolutional Network (GCN) and co-attention for aesthetic and emotion conjoint analysis. In our model, a stacked multimodal GCN network is proposed to encode the features under the guidance of the correlation matrix, and a co-attention module is designed to help the aesthetics and emotion feature representation learn from each other interactively. Experimental results indicate that our proposed model achieves competitive performance on the IAE dataset. Progressive results on the AVA and ArtPhoto datasets also prove the generalization ability of our model.

Keywords: image aesthetics and emotion analysis; multi-out learning; image semantic recognition; multimodal learning; Graph Convolutional Network; co-attention; intermediate supervision

Citation: Miao, H.; Zhang, Y.; Wang, D.; Feng, S. Multi-Output Learning Based on Multimodal GCN and Co-Attention for Image Aesthetics and Emotion Analysis. *Mathematics* **2021**, *9*, 1437. https://doi.org/10.3390/math9121437

Academic Editors: Darian M. Onchis, Dan Pescaru, Flavia Micota, Pedro Real and Codruta Istin

Received: 18 May 2021
Accepted: 17 June 2021
Published: 20 June 2021

1. Introduction

Current massive media and interactive modes have been developing since the Internet revolution. Presently, people tend to enjoy, share and store images of higher aesthetic quality in social networking [1]. At the same time, they often want to convey their emotional information to others by images [2]. Both aesthetics and emotion are the abstract and high-level semantic information of images, which contains lots of subjectivity [3]. There are a wide variety of applications in aesthetic assessment and emotion recognition, such as photo album management [4], on-line photo suggestion [5], image retrieval [6], digital monitoring [7], and news analysis [8]. Therefore, it is significant to analyze the aesthetics and emotion of images [9].

The image aesthetics study is typically cast into a classification or regression problem where the visual content is mapped to the aesthetic ratings provided by human annotators [10,11]. The emotion recognition of images is trying to predict the aroused human emotion when given a particular piece of visual content [12,13]. Many early works have mostly focused on manually crafted methods, which rely on various pixel-level features as well as feature representations such as color histograms [14],texture descriptors [15] and SIFT [16]. In recent years many deep neural network models, especially Convolutional

Neural Network (CNN), have achieved the state-of-the-art performance on many computer vision-related tasks [17–20]. An important reason for the success of CNN is that many large-scale datasets are available, such as AVA [21], MS_COCO [22], T4SA [23] and PhotoArt [24]. These datasets have tremendously promoted the development on many visual research areas, for example, image semantic segmentation [25,26], aesthetics assessment [27,28], and emotion recognition [29,30].

However, most studies treat image aesthetics assessment and emotion recognition as two independent studies, ignoring the fact that they are both people's mental responses to visual stimuli and correlated, i.e., an image can touch one's heart not only because of its visual content, but also its aesthetic composition. Images with higher aesthetic quality usually arouse people some positive emotions. Psychological studies have proved that human aesthetics as well as emotion can be aroused by visual content in the same time and affected with each other [31,32].

Besides that, the composition information of different visual elements and the semantic correlations between them define the harmony of an image [33]. Both aesthetics assessment and emotion recognition can benefit from capturing and exploring such important dependencies, which exist everywhere in our life, in visual and textual form [34,35]. Constructing a new large-scale multimodal dataset for it can be helpful but too costly. An alternative approach is to combine existing datasets with common large-scale corpora in some way.

Motivated by all of this, in this paper, we propose a multi-output learning model based on multimodal GCN and co-attention for aesthetics and emotion analysis. A summary of our work is as follows:

1. Assisted by target and scene recognition, we extract semantic labels from images. Via combining the co-occurrence frequency of the tags in the image dataset with their similarity in the external large-scale textual corpus, we construct two correlation matrices for aesthetics assessment and emotion recognition, respectively. Besides that, we also produce a 2D marked feature map to map the correlation similarity onto the image.
2. We convert an image into a regional graph in which each node denotes one region, and any two nodes are connected by a weighted edge. The weight is obtained by combining their visual similarity and correlation similarity. We perform reasoning on the graph with stacked graph convolution and acquire composition-aware re-encoded feature representation.
3. A co-attention module is designed to capture the mutual impacts between aesthetics and emotion for further enhancing the feature representations.

The main contributions of our work are as follows:

- We propose a novel end-to-end trainable multi-output learning framework for image aesthetics and emotion analysis.
- We extract semantic labels from images, construct multimodal correlation matrix, and propose a unique regional feature encoding mechanism that can capture potential aesthetic and emotional relationships.
- We present to use co-attention mechanism to jointly generate the aesthetics-guided and emotion-guided attention, which leads to better understanding of image aesthetics and emotion from each other interactively.
- We evaluate our method on three datasets, and our proposed method consistently achieves superior performance over previous approaches.

The remainder of this work is organized as follows: in Section 2, we discuss previous relevant works on image aesthetics assessment and emotion recognition. Section 3 details our proposed multi-output learning model based on multimodal GCN and co-attention for aesthetic and emotion analysis. In Section 4, we describe the datasets and implementation details in our experiments, and then discuss the experimental results. Finally, in Section 5, we conclude our work and present the future direction.

2. Related Work

In this section, we discuss the related work of our proposed method. Before the popularity of CNN, most emotion recognition studies has been dominated by traditional methods using manually crafted features or shallow classifiers such as local binary patterns (LBPs) [13,15,36], the Facial Action Coding System (FACS) with action units (AUs) [22,23], and sparse learning [13]. In the last few years, deep neural networks have contributed a lot to performance improvements concerning other popular learning algorithms, such as Naive Bayes and SVM [37]. In particular, Long Short-Term Memory Networks (LSTM) [38–40] and Convolutional Neural Networks (CNN) [29,41,42] were used for many emotion recognition tasks, such as Twitter [39] and some other image datasets [43]. Most existing visual sentiment classifiers have been trained on images, in which the emotion category is annotated by crowd-sourcing. Researchers propose deep networks to detect emotions by encoding visual features and map them with true labels.

As for aesthetics assessment, the estimation of image styles, aesthetics, and quality has been actively investigated over the past few decades. Similar to emotion recognition studies, the work in image aesthetics evaluation experience the process from traditional ways [21] to deep learning models [44,45]. K. Sheng et al. [46] presented an attention-based multi-patch aggregation network that enhances training signals by assigning relatively larger weights to misclassified image patches. Bowen Pan et al. [47] proposed a multi-task deep convolutional rating network to learn the aesthetic score and attributes simultaneously through a generative adversarial network.

However, researchers rarely focus on analyzing these two high-level and abstract concepts, image aesthetic quality and emotional expression, in an interactive and unified way, which has proved effective in the fields of computer [3,9,48] and psychology [31,32]. Yu et al. [49] extend a large-scale emotion dataset by further rating the aesthetic scores by volunteers and propose a hybrid multi-task learning model on unified aesthetics and emotion prediction. Nevertheless, they just concatenate the final output vectors which may lead to inadequate learning.

In view of the significant progress in cross-modal learning that the co-attention mechanism has made [50,51], we propose a co-attention model to encode aesthetics and emotion features in an interactive way simultaneously.

Besides that, most attention to previous work is paid to modeling the image itself to improve the performance, ignoring to explore intrinsic and high-level properties, such as the composition of semantic elements. Recently, GCN has been widely used for natural language processing and text emotion recognition [52,53], due to its powerful capacity of exploring relationships among semantic and emotional elements. GCN can focus on graph-structure data and encode features based on only themselves but also their relationships with each other. Some researchers have applied GCN to semantic vision, as CNN cannot reason rich dependency of different visual contents. Chen et al. [33] proposes a multi-label image classification model based on GCN. They build a directed graph over the object labels and map the label graph into a set of object classifiers. Liu et al. [34] takes the image as a graph of regional nodes and computes aesthetic properties by reasoning on the graph. Despite this, they only consider relationships that exist in their dataset. Combining textual contents to facilitate visual understanding is usually a very effective approach, but one of the limitations is the lack of multimodal data.

To address this issue, we transfer two large-scale textual corpora and construct the relation graph assisted by mining knowledge from them. In this paper, we present a multi-output learning model integrating GCN and co-attention to predict aesthetics and emotion in the images.

3. Proposed Model

In this section, we propose a model for image aesthetics and emotion transactional analysis assisted by multimodal GCN and co-attention mechanism. The pipeline of our proposed framework is shown in Figure 1. First, we conduct textual multi-label semantic

recognition and construct a correlation matrix for these labels assisted by external knowledge transferring. Taking an image as the input, we extract the regional feature map from the deep network and carry out information propagation on this regional graph via GCN. After that, we acquire a processed regional feature map. We train two parallel branches for aesthetics assessment and emotion recognition, respectively, and then send the two elaborate regional feature maps to the co-attention module. The co-attention module generates two attention mask matrix: one for the aesthetics feature and the other for the emotion feature. Finally, we gain two attentional feature descriptors and take them to perform prediction on each of those tasks.

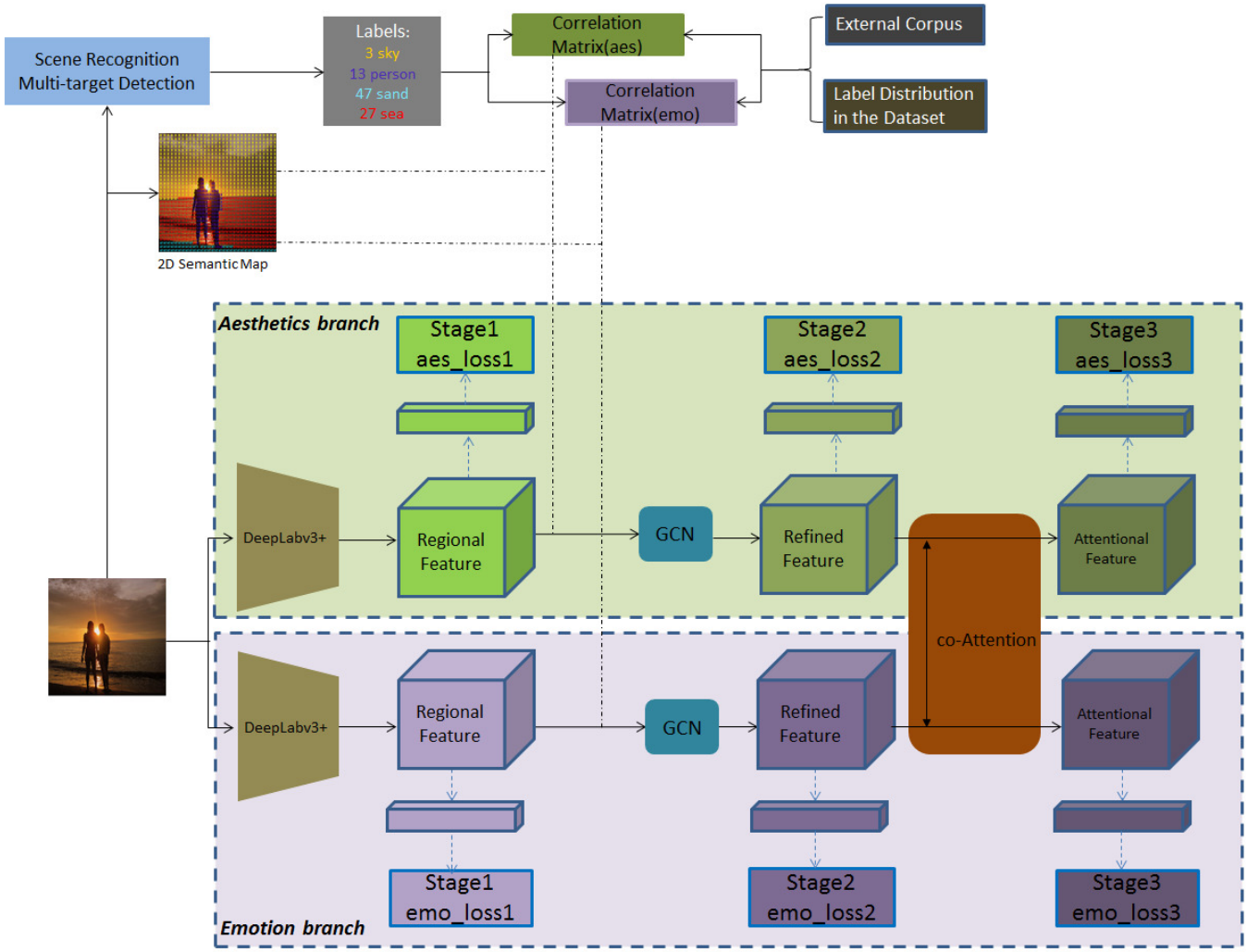

Figure 1. Illustration of our proposed framework.

3.1. Textual Multi-Label Semantic Recognition

Image aesthetics assessment and emotion recognition could profit from multimodal contents as they can provide more vivid and adequate information. For that purpose, we would like to extract multiple textual properties from each input image. We send the input image into a DeepLabv3+ [54] model that has already shown the state-of-art performance in extensive visual Semantic Segmentation, multi-target detection, and scene recognition research. As shown in Figure 2, corresponding labels for targets (with a confidence ≥ 0.7) and scenes in each image are acquired simultaneously from the DeepLabv3+ pre-trained on the MS_COCO [22] (for targets) and Ade20k dataset [55] (for targets and scenes). Thus, we obtain a group of semantic labels on the entire dataset, i.e., $L = \{L_1, L_2, \cdots, L_c\}$, where c denotes the number of categories.

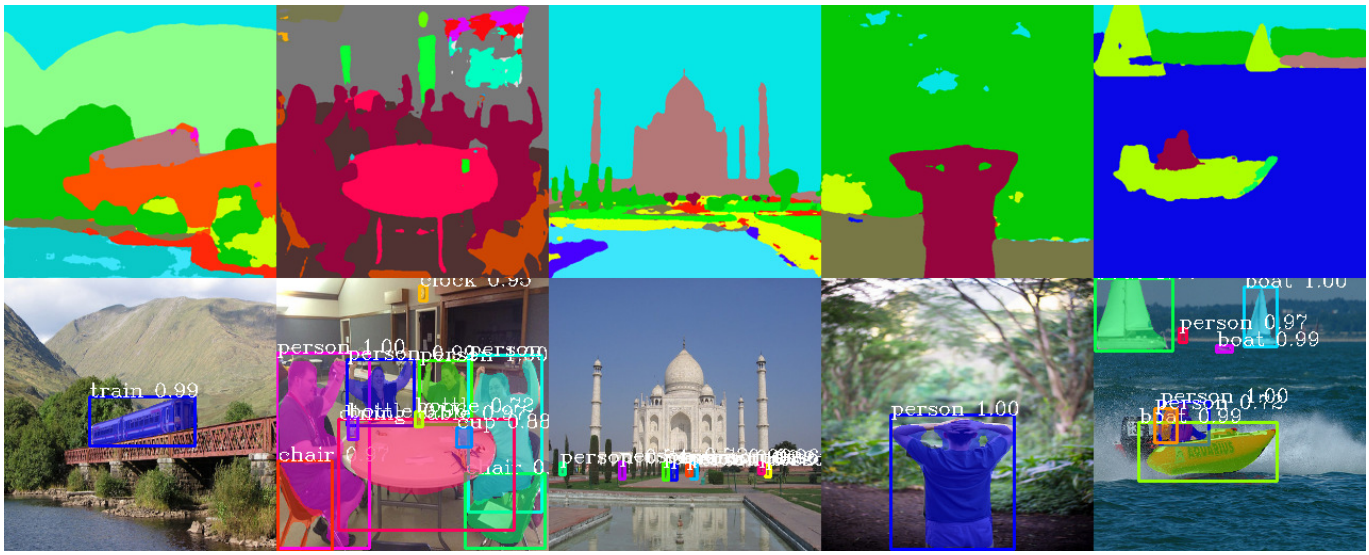

Figure 2. Examples of Multi-label Semantic Recognition. Targets and scenes in images are displayed in different colors, and targets with a confidence ≥ 0.7 are also shown.

3.2. Correlation Matrix for Textual Semantic Labels

For better using textual semantic attributes in images, we transfer external textual correlation knowledge to construct a correlation matrix to guide information propagation among the regional nodes in GCN. For aesthetics assessment, we generate a dataset that contains all user comments crawled from the provided page links in the AVA dataset [21] on https://www.dpchallenge.com (accessed on 19 June 2021) [56]. Each image is commented and scored by users, commenters, participants, and non-participants on the website. Figure 3 shows some examples in the AVA-comment dataset, where semantic labels are marked in red. For emotion recognition, we use the Twitter sentiment dataset [23]. With the training of these two corpora, we transform each semantic label into the task-specific word vector v^t, where t is the label of the aesthetics and emotion prediction task, via the GloVe word embeddings [57] model, severally.

To construct the correlation matrix, we define the correlation between the semantic labels via mining their co-occurrence patterns within the image dataset and model the semantic label correlation dependency for different task t in the form of C_{ij}^t :

$$C_{ij}^t = \left(\frac{n_{ij}^t}{n_j^t} \log \frac{N^t}{n_i^t} \right) s\left(v_i^t, v_j^t \right) \tag{1}$$

where n_{ij}^t means the concurring times of label L_i and label L_j in the same image, n_i^t or n_j^t denotes the number of images that contain label L_i or L_j, N^t is the total number of images in the dataset, $s\left(v_i^t, v_j^t \right)$ is the cosine similarity of semantic word vector v_i^t and v_j^t. The formulation of the cosine similarity is expressed in the form:

$$s\left(v_i^t, v_j^t \right) = \frac{v_i^t v_j^t}{\left\| v_i^t \right\| \left\| v_j^t \right\|} \tag{2}$$

The correlation matrix is unsymmetrical, as C_{ij}^t is not equal to C_{ji}^t. After that, we perform $L2$-normalization on C_{ij}^t and obtain $\hat{C}_{ij}^t$ as follows:

$$\hat{C}_{ij}^t = \frac{C_{ij}^t}{\| C^t \|_2} \tag{3}$$

Finally, we construct the correlation matrix $\hat{C}^t \in \mathbb{R}^{N_L \times N_L}$ for each of two prediction tasks, where N_L denotes the number of textual semantic labels.

Figure 3. Examples in the AVA-comment dataset. Semantic labels in comments are marked in red.

3.3. Visual Regional Feature Extraction

In this part, we view an image as a graph composed of local regions and would like to complete reasoning over the regional graph to promote image aesthetics assessment and emotion recognition. For this purpose, we take the image as the input and send it to the DeepLabv3+ to produce the 3D feature map in which each spatial region represents a local area of the image, and the arrangement of all regional feature maps describes a spatial composition of various visual components of an image. Besides that, we also produce a 2D semantic map in which each region is marked by an integer corresponding to one of textual semantic labels. For visualization, we show the marked map on the original image in Figure 4, where semantic labels are displayed as colored numbers, for example, 3 is Sky, 5 is Tree, 13 is Person, 17 is Mountain, 27 is Sea, 104 is Ship, and so on. The 3D feature map has the same size $(h \times w)$ as the 2D semantic map, and that is the latter is an additional label group for the former.

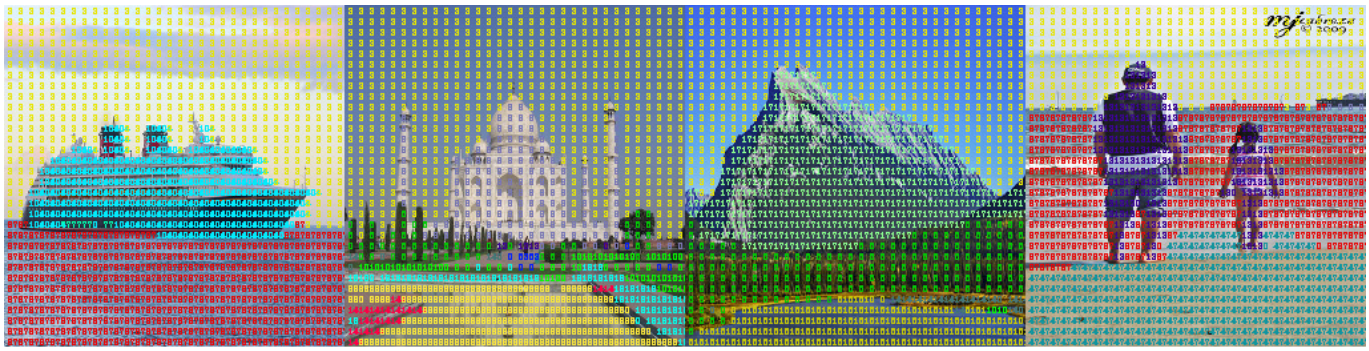

Figure 4. Examples of 2D semantic map visualization. Semantic labels are marked as numbers in different colors.

3.4. GCN for Visual Regional Feature Encoding

To capture correlations between visual image contents and explore these interdependencies, we construct a multimodal regional correlation graph guiding composition-aware feature encoding. In this part, we adopt graph convolution operation under the interactive guidance of textual semantic properties as the regional dependency modeling mechanism to learn a more competitive feature representation for image aesthetics assessment and emotion recognition.

In the beginning, given the feature map of dimensions $h \times w \times d$, we denote a set of m regions as $R = \{r_1, r_2, \cdots, r_m\}$, where $m = h \times w$ is the total number of regional feature vectors. The graph structure is composed of multiple node vectors. In other words, to construct the graph, we represent each node as the local region, which is a d-dimensional vector. In addition, we connect each pair of nodes with a weighted edge. Considering multimodal content, we calculate the weight A_{ij} as follows:

$$A_{ij} = \hat{C}_{\tilde{i}\tilde{j}} \hat{V}_{ij} \tag{4}$$

where $\hat{C}$ is the correlation matrix given in Equation (3) , $\tilde{i}$ and $\tilde{j}$ are the corresponding label of i and j in the 2D semantic map, and $\hat{V}_{ij}$ is the visual similarity normalized by $softmax$ like [], represented using following equations:

$$V_{ij} = (Ur_i)^T (Vr_j) \tag{5}$$

$$\hat{V}_{ij} = \frac{exp(V_{ij})}{\sum_{k=1}^{m} exp(V_{ik})} \tag{6}$$

where $U \in \mathbb{R}^{d \times d}$ and $V \in \mathbb{R}^{d \times d}$ are two transformation matrices [58] that will be constantly updated with the network training.

For the weight matrix A_{ij}, we filter out those pretty small values by a threshold of 0.1 and obtain a new weight matrix $\hat{A}$ as:

$$\hat{A}_{ij} = \begin{cases} \hat{A}_{ij} & if \quad \hat{A}_{ij} \geq 0.1 \\ 0 & if \quad \hat{A}_{ij} < 0.1 \end{cases} \tag{7}$$

As an adjacency matrix of the relationship between nodes, $\hat{A} \in \mathbb{R}^{m \times m}$ describes the interdependence of regional areas in the image. In the graph reasoning stage, we use multi-layer GCN to transmit the structural information of nodes through the adjacent matrix as follows:

$$H^{l+1} = ReLu\left(\hat{A} H^l B^l\right) \tag{8}$$

where H^l is the lth layer updated node feature, $B^l \in \mathbb{R}^{d \times d}$ is lth layer learnable transformation matrix, and $ReLu$ is the activation function. Figure 5 is a visualization of multi-layer GCN, where regional nodes with different semantic labels are marked with different colors, and edges denote the relationships between nodes with different multimodal contents. During the stacked graph reasoning, each node feature is updated only by itself and the nodes that are correlated with multimodal content.

3.5. Co-Attention Module for Aesthetics and Emotion Analysis

In this part, we model the mutual interaction between regional image features of different tasks based on the collaborative attention mechanism. Specifically, given the task-specific visual feature representation $X^t \in \mathbb{R}^{m \times d}$, we construct the attention matrix and generate two attentional feature representations for aesthetics and emotion.

First, similar to [59], we construct the attention matrix $M \in \mathbb{R}^{m \times m}$ that calculates the similarity between X^{aes} and X^{emo} via the measurement function:

$$M = \tanh\left(X^{aes} W_s X^{emoT} + b_s\right) \tag{9}$$

where $W_s \in \mathbb{R}^{d\times d}$ and $b_s \in \mathbb{R}^{m\times m}$ are both learnable matrices, and tanh is the activation function.

With the affinity matrix M, two task-specific attentional feature representations $\tilde{X}^t \in \mathbb{R}^{m\times k}$ can be calculated as follows:

$$\tilde{X}^{aes} = \tanh\left(X^{aes}W_a + \left(X^{aes}W_a \bigotimes MX^{emo}W_e\right) + b_e\right) \tag{10}$$

$$\tilde{X}^{emo} = \tanh\left(X^{emo}W_e + \left(X^{emo}W_e \bigotimes M^T X^{aes}W_a\right) + b_a\right) \tag{11}$$

where $W_a \in \mathbb{R}^{d\times k}$, $W_e \in \mathbb{R}^{d\times k}$, $b_a \in \mathbb{R}^{m\times k}$, and $b_e \in \mathbb{R}^{m\times k}$ are all learnable matrices, and $\bigotimes$ denotes element-wise product.

Finally, we feed the attentional feature into a three-layer network consisting of a convolutional layer and two FC layers to obtain a 1D vector for aesthetic and emotional evaluation.

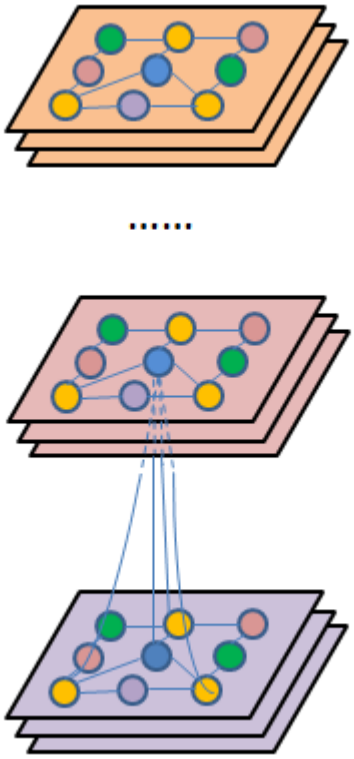

Figure 5. Visualization of multi-layer GCN, a framework for stacked graph reasoning with the regional map based on relationships between nodes. Nodes with different semantic labels are marked with different colors, and edges denote multimodal relationships between nodes. Each node feature is gradually updated only by itself and the nodes with edges connected to it.

3.6. Intermediate Supervision

Our model architecture contains three stages: (1) learning image features via deep networks; (2) graph reasoning with the regional map; and (3) collaborative attention encoding. To avoid the gradient vanishing with repetitious backpropagation, we apply intermediate supervision, which has demonstrated strong performance among the methods of multiple iterative stages [60,61], to produce a loss on each stage. We feed the feature into a three-layer network consisting of a convolutional layer and two FC layers to obtain a 1D vector and use the cross-entropy loss as the objective function. It is expressed as follows:

$$Loss^t = Loss_1^t + Loss_2^t + Loss_3^t \tag{12}$$

$$Loss_{total} = Loss^{aes} + Loss^{emo} \tag{13}$$

4. Experiments and Results

In this section, we evaluate our proposed multi-output learning model integrating multimodal GCN and co-attention. First, we introduce the datasets and the experimental settings in our work. Then the results of our proposed method are compared against several baselines. Finally, we discuss the performance of our model.

4.1. Datasets

We conduct experiments on the Images with Aesthetics and Emotion (IAE) dataset [49], a new dataset associated with both aesthetic and emotional labels. Specifically, IAE is an extension of the earlier work in [43], where more than 22,000 images are manually divided into eight emotion categories, i.e., amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. Each category consists of more than 1100 images. These images are next rated with ten volunteers from the aesthetic perspective in [49]. Thus, the IAE dataset contains 11,550 high-aesthetic and 10,536 low-aesthetic images. Figure 6 shows the distribution of 8-category emotion in high-aesthetic and low-aesthetic images. As we can see, most positive emotions, especially awe and contentment, appear more frequently with high aesthetics, while negative emotions always come with low aesthetics. It provides empirical support that aesthetics and emotion are correlated and can be studied interactively. For a fair comparison, we follow the same data partition used in their work, i.e., 70% of images for training, 10% of images for validation, and the rest for testing.

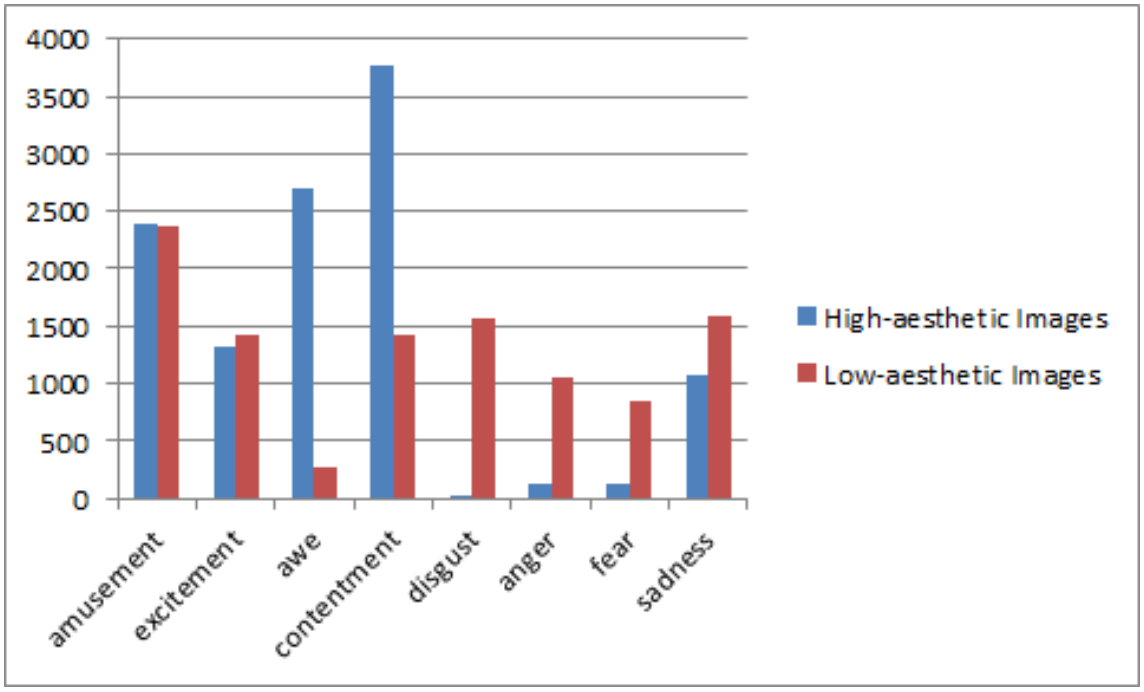

Figure 6. Data distribution in IAE dataset. The first four emotions (i.e., amusement, excitement, awe, and contentment) are positive, while the last four (i.e., disgust, anger, fear, and sadness) are negative.

To verify the generalization ability of the single prediction task for aesthetics and emotion, we also choose two benchmark datasets, AVA [21] and ArtPhoto [24]. AVA is the largest publicly available dataset for aesthetic visual analysis and contains scored images collected during the digital photography contest on https://www.dpchallenge.com (accessed on 19 June 2021). ArtPhoto is a set of artistic photographs downloaded from an art sharing site, https://www.deviantart.com (accessed on 19 June 2021) [62], taking the emotion categories as search terms.

To capture interdependencies between visual image contents, we construct a correlation matrix that helps to guide composition-aware feature encoding by transferring external textual knowledge. For aesthetics assessment, we crawl the user comments for

images in AVA from https://www.dpchallenge.com (accessed on 19 June 2021) to form the AVA-Comments dataset. All quotes and extra HTML tags such as links are removed. As for emotion recognition, we use the Twitter sentiment dataset [23] in which more than 3 million tweets containing both text and images are collected. Following their work, we select only the tweets classified with a confidence ≥ 0.85 and marked as positive or negative, thus get more than 550K tweets. With the training of these two corpora, we transform each semantic label into the task-specific word vector via the GloVe word embeddings model, severally.

4.2. Experimental Settings

DeepLabv3+ network with Xception [63] as its backbone is used for our feature encoding. The input raw image is resized to 224×224, and the output feature map is $28 \times 28 \times 256$ after the Atrous Spatial Pyramid Pooling (ASPP) [64] module, followed by a 1×1 convolution layer. The number of GCN layers is three, and the dimension k of feature space in the co-attention module is 128. In our networks, the model is initialized on ImageNet and then fine-tuned end-to-end with images for aesthetics and emotion multi-output prediction. For network optimization, SGD is used as an optimizer. The base learning rate is 10^{-4} and reduced by a factor of 10 every ten epochs. The momentum, weight decay, total epoch and batch size are set to 0.9, 10^{-5}, 50 and 32, respectively. Our networks are implemented based on PyTorch.

4.3. Comparison Evaluation with Baselines

In this section, we first present a comparison evaluation on IAE to demonstrate the effectiveness of our method. Then we evaluate the performance of our model in the single prediction task for aesthetics and emotion on AVA and ArtPhoto, respectively.

Table 1 shows the performance of our method on the IAE dataset along with the different results of others. ResNet50 [65], WRN [66], MLSP [67], RGNet [34], CycleEmotionGAN [68], and APSE [69] are single-task learning models. SSNet [70], CSNet [71], AENet-FL and AENet [49] are multi-task learning models. The experimental results in Table 1 reveal that our proposed model outperforms other methods on various networks. Compared with single-task learning methods, our method outperforms two base networks ResNet50 and WRN by 9.78% and 9.7% for aesthetics assessment. However, for emotion recognition, the improvement is 9.05% and 7.49%. We also introduce four single-task learning models as baselines, of which MLSP and RGNet are for aesthetic assessment, while CycleEmotionGAN and APSE are for emotion recognition. Against them, our method exhibits at least 5.18% and 3.95% performance improvement in the respective prediction tasks. Besides that, it is interesting to see that the best result of previous multi-task learning models (AENet) also beats those single-task learning models, which give a proof of the aesthetic and emotion conjoint analysis. For multi-task learning models, such as SSNet, our method achieves 6.84% and 9.91% higher accuracy for aesthetics and emotion tasks, respectively. Compared with CSNet, our method outperforms it by 6.93% and 6.97% for aesthetics and emotion. AENet is an extension of AENet-FL, as the former adds the extra fusion layers and achieves 81.05% and 66.23% accuracy, overcoming the latter. In comparison, our method achieves 85.63% and 70.14%, surpassing them by a large margin of at least 4.58% and 3.91% for aesthetics and emotion, severally. The competitive results demonstrate the effectiveness of our proposed model, both in aesthetics assessment and emotion recognition.

As mentioned above, to verify the generalization ability of the single prediction task for aesthetics and emotion, we also train our model on IAE and test it on AVA and ArtPhoto. The comparison results can be found in Table 2. As we can see, there is a sharp decline in performance for all methods. It is due to the heterogeneity between the different data sets. In comparison, our method achieves 80.2% and 40.77% accuracy, surpassing them by a large margin for both prediction tasks, especially for emotion prediction. It may be because we pre-extract semantic labels from images in the test set and construct a correlation matrix based on their co-occurrence frequency and textual similarity. With the help of transferring

the correlation, the information gap could be alleviated. The competitive results prove that the multimodal GCN module based on external knowledge transfer does improve the prediction performance and generalization ability for both aesthetics assessment and emotion recognition.

Table 1. Comparison on IAE for aesthetics assessment and emotion recognition, respectively (aesthetics is abbreviated as Aes, emotion as Emo, and accuracy as ACC).

Method	Description %	IAE Aes ACC %	IAE Emo ACC %
ResNet50	Single-task for Aes and Emo respectively.	74.85	61.09
WRN	Single-task for Aes and Emo respectively.	75.16	62.65
MLSP	Single-task for only Aes.	78.68	—
RGNet	Single-task for only Aes.	80.45	—
CycleEmotionGAN	Single-task for only Emo.	—	65.78
APSE	Single-task for only Emo.	—	66.19
SSNet	Multi-task for both Aes and Emo.	77.79	60.23
CSNet	Multi-task for both Aes and Emo.	77.70	63.17
AENet-FL	Multi-task for both Aes and Emo.	76.68	61.46
AENet	Multi-task for both Aes and Emo.	81.05	66.23
Ours	Multi-task for both Aes and Emo.	**85.63**	**70.14**

Table 2. Comparison on AVA for aesthetics assessment and on ArtPhoto for emotion recognition, respectively (aesthetics is abbreviated as Aes, emotion as Emo, and accuracy as ACC).

Method	Description %	IAE Aes ACC %	IAE Emo ACC %
ResNet50	Single-task for Aes and Emo respectively.	71.18	24.83
WRN	Single-task for Aes and Emo respectively.	71.33	27.02
MLSP	Single-task for only Aes.	71.89	—
RGNet	Single-task for only Aes.	73.47	—
CycleEmotionGAN	Single-task for only Emo.	—	30.83
APSE	Single-task for only Emo.	—	29.54
SSNet	Multi-task for both Aes and Emo.	68.39	23.86
CSNet	Multi-task for both Aes and Emo.	70.48	27.30
AENet-FL	Multi-task for both Aes and Emo.	70.29	24.22
AENet	Multi-task for both Aes and Emo.	72.83	27.92
Ours	Multi-task for both Aes and Emo.	**80.2**	**40.77**

4.4. Discussion on the Performance and Results of Our Model

Figure 7 demonstrates the normalized confusion matrix of 8-category emotions on the IAE and ArtPhoto datasets to analyze the performance of our proposed method in each emotion category. They both contain eight emotion categories, i.e., amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. In the IAE dataset, the Anger and Fear categories have lower accuracy than others. Many Anger and Fear categories are misclassified as sadness categories. Besides that, the excitement category can be easily mistaken for the amusement category, which may be attributed to the similar emotion features of the two categories. Moreover, when we test on the ArtPhoto, our model does not perform as well as before. Only the Fear category can be correctly distinguished with a probability of more than 50%.

Figure 8 is a visualization of intermediate supervision, as mentioned in Section 3.6. As we can see, in both aesthetics and emotion branches, the loss in all three stages is decreasing as the training continues. It proves that the gradient is back-propagated to the whole end-to-end network, and all parameters are updated.

Figure 9 summarizes the results in different stages on IAE, as mentioned in Section 3.6. For aesthetics assessment and emotion recognition, both the GCN module and co-attention

module in our model can boost performance. Moreover, it is interesting to find out that our co-attention module performs better in enhancing the prediction of the Awe and Disgust categories, which may be attributed to the similar consistency and correlation between the two emotion categories and aesthetic perception.

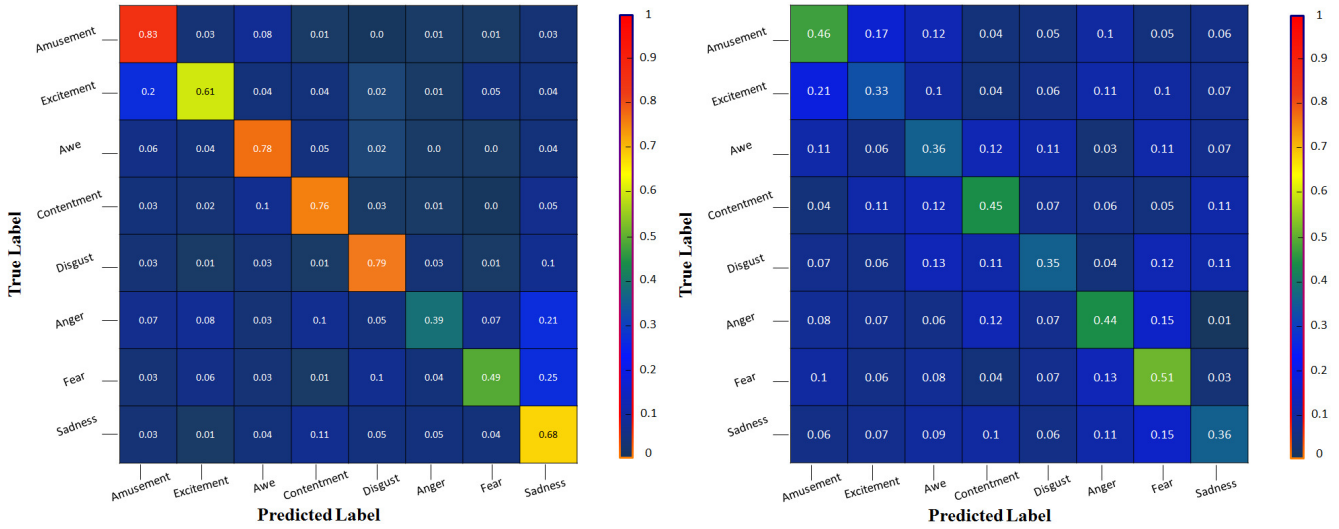

Figure 7. Normalized confusion matrix of 8-category emotion on the IAE and ArtPhoto datasets.

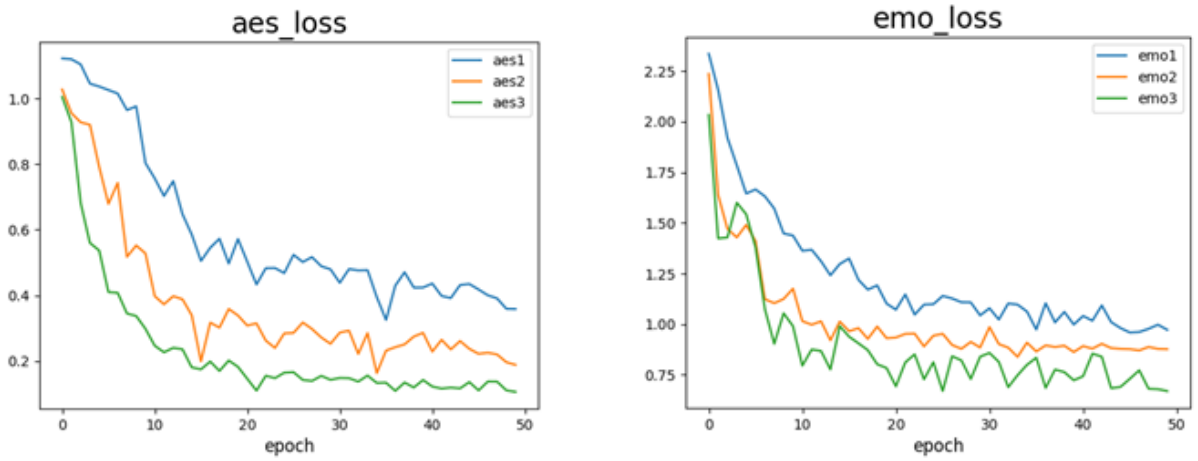

Figure 8. Visualization of intermediate supervision.

We conduct several experiments to evaluate the performance of the GCN module. We have designed 5 variants of our model, without or with different GCN. The first variant is that we remove the GCN module and send features to the co-attention module directly. The second variant is that we set all values in the weight matrix $\hat{A}$ to 1, which turns the GCN into the fully connected networks (FCN). For the rest of the variants, we take $\hat{C}$, $\hat{V}$, and $\hat{A} = \hat{C} \otimes \hat{V}$ ($\otimes$ denotes element-wise product) as the textual, visual, and multimodal weight matrix for GCN. The performance of these models is shown in Table 3. For a fair comparison, we use the same experimental settings among all the variants. From the results, we can see that the variant with FCN does not perform better than the variant without GCN. Both perform worse than the rest of the variants with different modality GCN, which means that simply increasing the number of layers does not work, and GCN does explore the dependencies among visual elements. As for single modality GCN, visual and textual correlation reasoning achieve a similar accuracy on both aesthetics and emotion prediction tasks. Both are defeated by multimodal GCN, by at least 1.67 % and 1.33 % higher accuracy on aesthetics and emotion tasks, which provides evidence for the validity of our multimodal GCN module.

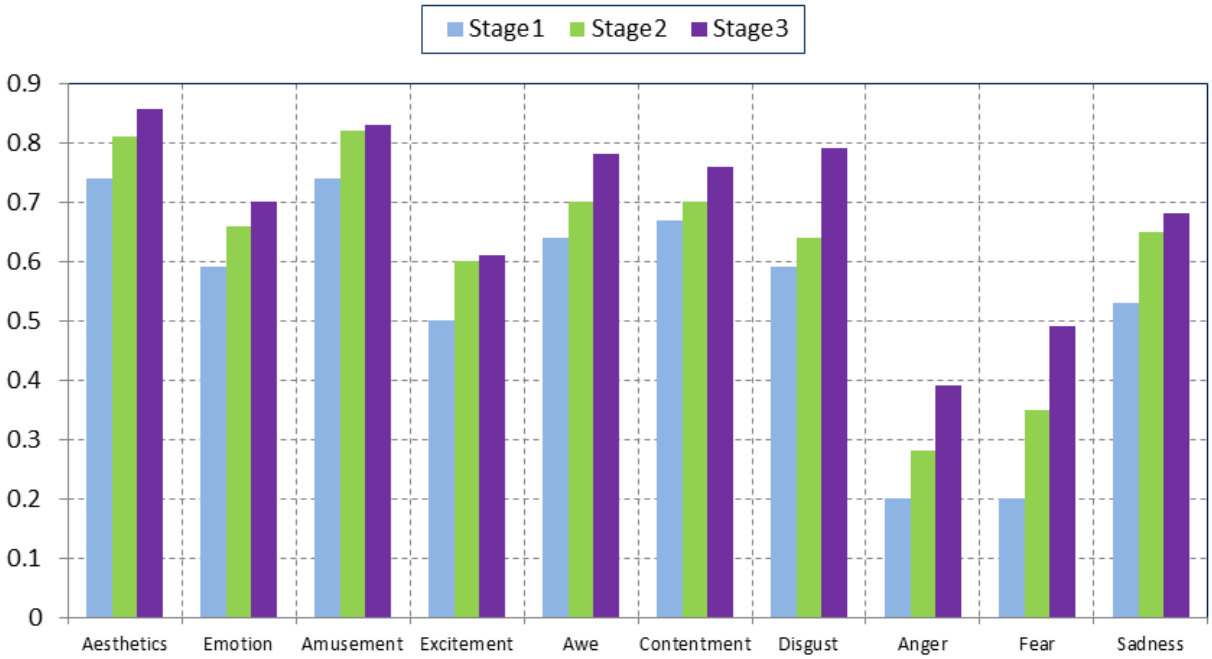

Figure 9. Results in different stages on IAE.

Table 3. Comparison of the model without or with different GCN on IAE for aesthetics assessment and emotion recognition, respectively (aesthetics is abbreviated as Aes, emotion as Emo, and accuracy as ACC).

Variants	IAE Aes ACC %	IAE Emo ACC %
without GCN	80.48	65.63
with FCN	80.55	65.89
with textual GCN	83.77	68.81
with visual GCN	83.96	68.35
Multimodal GCN	**85.63**	**70.14**

5. Conclusions

Every day, more and more humans share their experiences, communicate their moods and state of mind with images on social networks. Meanwhile, they always look forward to enjoying higher quality images, which are more likely to trigger their emotion. Therefore, aesthetics assessment and emotion recognition are two fundamental problems in user perception understanding. Although the two tasks are correlated and mutually beneficial, they are usually solved separately in existing studies.

For image aesthetics and emotion analysis, they belong to image processing and pattern recognition in computational mathematics. Methods of machine learning and deep learning have provided tremendous assistance in mathematical problem-solving in recent years. Using the deep neural network, we map images and abstract concepts into numerical matrices. In addition, after a series of mathematical operations, we build relationships between visual data with labels in aesthetic and emotional perspectives.

In this paper, we propose an end-to-end multi-output deep learning model based on multimodal GCN and co-attention for image aesthetics and emotion conjoint analysis. We extract semantic labels from the image and construct a correlation matrix, assisted by the label distribution in the dataset and external textual information by corpus-based knowledge transferring. As the correlation matrix mapped onto the image, we perform stacked graph reasoning on the regional image map and then obtain a composition-aware re-encoded feature representation. After that, we send these features into the co-attention module and let them learn from each other interactively and selectively. To avoid the

gradient vanishing with repetitious backpropagation, we apply intermediate supervision to produce a loss on each stage.

Experimental results on the multiple datasets demonstrate the performance of our proposed method for aesthetics and emotion analysis. On the Image with Aesthetics and Emotion (IAE) dataset, our method achieves 85.63% and 70.14% accuracy for aesthetics and emotion tasks, respectively. Compared with the baselines, our method outperforms both single- and multi-task learning models by at least 4.58% and 3.91% accuracy. We also test on AVA and ArtPhoto by performing a cross-dataset evaluation with IAE to verify the generalization ability of our method for two prediction tasks. Our method achieves 80.2% and 40.77% accuracy, surpassing the baselines by a large margin for both two tasks. However, the result for emotion prediction is needed to be further improved. In addition, the visualized results and the comparison results show that the intermediate supervision, GCN module, and co-attention module can boost the performance of our model.

As future work, we plan to introduce some more complex and nuanced relationships, such as camera style and emotion cause, to help the model simulate high-level visual perceptions.

Author Contributions: Methodology, software, validation, writing—original draft preparation, visualization, investigation, data curation, H.M.; conceptualization, formal analysis, writing—review and editing, supervision, resources, D.W., Y.Z. and S.F.; project administration, funding acquisition, D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China under grant 2018YFB1004700, and National Natural Science Foundation of China (61772122, 61872074).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors are grateful to Jun Yu and Lucia Vadicamo for providing the dataset used in the experiments.

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

References

1. Chen, Y.; Klopp, J.; Sun, M.; Chien, S.; Ma, K. Learning to Compose with Professional Photographs on the Web. In Proceedings of the 25th ACM International Conference on Multimedia (MM), Mountain View, CA, USA, 23–27 October 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 37–45.
2. You, Q.; Luo, J.; Jin, H.; Yang, J. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the 29th Association-for-the-Advancement-of-Artificial-Intelligence (AAAI) Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; The AAAI Press: Palo Alto, CA, USA, 2015; pp. 381–388.
3. Datta, R.; Li, J.; Wang, J.Z. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In Proceedings of the 15th IEEE International Conference on Image Processing (ICIP), San Diego, CA, USA, 12–15 October 2008; IEEE Computer Society: Washington, DC, USA, 2008; pp. 105–108.
4. Deng, X.; Cui, C.; Fang, H.; Nie, X.; Yin, Y. Personalized Image Aesthetics Assessment. In Proceedings of the 25th ACM Conference on Information and Knowledge Management (CIKM), Singapore, 6–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 2043–2046.
5. Zhao, S.; Zhao, X.; Ding, G.; Keutzer, K. EmotionGAN: Unsupervised Domain Adaptation for Learning Discrete Probability Distributions of Image Emotions. In Proceeding of 26th ACM International Conference on Multimedia (MM), Seoul, Korea, 22–26 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1319–1327.
6. Cui, C.; Fang, H.; Deng, X.; Nie, X.; Dai, H.; Yin, Y. Distribution-oriented Aesthetics Assessment for Image Search. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, 7–11 August 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1013–1016.
7. Fan, Y.; Lam, J.C.; Li, V.O. Video-based Emotion Recognition Using Deeply-Supervised Neural Networks. In Proceedings of the International Conference on Multimodal Interaction (ICMI), Boulder, CO, USA, 16–20 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 584–588.

8. Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; Zhou, G. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In Proceedings of the 29-th International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 10–16 August 2019; IJCAI: San Mateo, CA, USA, 2019; pp. 5415–5421.
9. Kostoulas, T.; Chanel, G.; Muszynski, M.; Lombardo, P.; Pun, T. Films, Affective Computing and Aesthetic Experience: Identifying Emotional and Aesthetic Highlights from Multimodal Signals in a Social Setting. *Front. ICT* **2017**, *4*, 11. [CrossRef]
10. Kong, S.; Shen, X.; Lin, Z.L.; Mech, R.; Fowlkes, C. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: New York, NY, USA, 2016; pp. 662–679.
11. Lu, X.; Lin, Z.; Jin, H.; Yang, J.; Wang, J.Z. RAPID: Rating Pictorial Aesthetics using Deep Learning. In Proceedings of the ACM International Conference on Multimedia (MM), Orlando, FL, USA, 3–7 November 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 457–466.
12. Lee, J.; Kim, S.; Kim, S.; Park, J.; Sohn, K. Context-Aware Emotion Recognition Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; IEEE Computer Society: Washington, DC, USA, 2019; pp. 10142–10151.
13. Yu, Z.; Zhang, C. Image based Static Facial Expression Recognition with Multiple Deep Network Learning. In Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), Seattle, WA, USA, 9–13 November 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 435–442.
14. Shan, C.; Gong, S.; McOwan, P.W. Facial expression recognition based on Local Binary Patterns: A comprehensive study. *Image Vis. Comput.* **2009**, *27*, 803–816. [CrossRef]
15. Zhao, G.; Pietikainen, M. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef] [PubMed]
16. Zhong, L.; Liu, Q.; Yang, P.; Huang, J.; Metaxas, D.N. Learning Multiscale Active Facial Patches for Expression Analysis. *IEEE Trans. Cybern.* **2015**, *45*, 1499–1510. [CrossRef] [PubMed]
17. Joshi, M.R.; Nkenyereye, L.; Joshi, G.P.; Islam, S.M.R.; Abdullah-Al-Wadud, M.; Shrestha, S. Auto-Colorization of Historical Images Using Deep Convolutional Neural Networks. *Mathematics* **2020**, *8*, 2258. [CrossRef]
18. Zhou, Z.; Wang, M.; Cao, Y.; Su, Y. CNN Feature-Based Image Copy Detection with Contextual Hash Embedding. *Mathematics* **2020**, *8*, 1172. [CrossRef]
19. Liu, F.; Zhou, X.; Yan, X.; Lu, Y.; Wang, S. Image Steganalysis via Diverse Filters and Squeeze-and-Excitation Convolutional Neural Network. *Mathematics* **2021**, *9*, 189. [CrossRef]
20. Darabant, A.S.; Borza, D.; Danescu, R. Recognizing Human Races through Machine Learning-A Multi-Network, Multi-Features Study. *Mathematics* **2021**, *9*, 195. [CrossRef]
21. Murray, N.; Marchesotti, L.; Perronnin, F. AVA: A large-scale database for aesthetic visual analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; IEEE Computer Society: Washington, DC, USA, 2012; pp. 2408–2415.
22. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: New York, NY, USA, 2014; pp. 740–755.
23. Vadicamo, L.; Carrara, F.; Cimino, A.; Cresci, S.; Dell'Orletta, F.; Falchi, F.; Tesconi, M. Cross-Media Learning for Image Sentiment Analysis in the Wild. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Venice, Italy, 22–29 October 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 308–317.
24. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th International Conference on Multimedia (MM), Firenze, Italy, 25–29 October 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 83–92.
25. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
26. Teso-Fz-Betoño, D.; Zulueta, E.; Sánchez-Chica, A.; Unai, F.; Aitor, S. Semantic Segmentation to Develop an Indoor Navigation System for an Autonomous Mobile Robot. *Mathematics* **2020**, *8*, 855. [CrossRef]
27. Deng, Y.; Loy, C.C.; Tang, X. Aesthetic-Driven Image Enhancement by Adversarial Learning. In Proceedings of the ACM Multimedia Conference on Multimedia Conference (MM), Seoul, Korea, 22–26 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 870–878.
28. Sheng, K.; Dong, W.; Chai, M.; Wang, G.; Zhou, P.; Huang, F.; Hu, B.; Ji, R.; Ma, C. Revisiting Image Aesthetic Assessment via Self-Supervised Feature Learning. In Proceedings of the 30-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; The AAAI Press: Palo Alto, CA, USA, 2020; pp. 5709–5716.
29. Campos, V.; Jou, B.; Giró-i-Nieto, X. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image Vis. Comput.* **2017**, *65*, 15–22. [CrossRef]
30. Kao, Y.; He, R.; Huang, K. Deep Aesthetic Quality Assessment with Semantic Information. *IEEE Trans. Image Process.* **2017**, *26*, 1482–1495. [CrossRef] [PubMed]
31. Kätsyri, J.; Ravaja, N.; Salminen, M. Aesthetic images modulate emotional responses to reading news messages on a small screen: A psychophysiological investigation. *Int. J. Hum. Comput.* **2012**, *70*, 72–87. [CrossRef]

32. Leder, H.; Belke, B.; Oeberst, A.; Augustin, D. A model of aesthetic appreciation and aesthetic judgments. *Br. J. Psychol.* **2004**, *95*, 489–508. [CrossRef]
33. Chen, Z.; Wei, X.; Wang, P.; Guo, Y. Multi-Label Image Recognition with Graph Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; IEEE Computer Society: Washington, DC, USA, 2019; pp. 5177–5186.
34. Liu, D.; Puri, R.; Kamath, N.; Bhattacharya, S. Composition-Aware Image Aesthetics Assessment. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; IEEE Computer Society: Washington, DC, USA, 2020; pp. 3558–3567.
35. Gao, Q.; Zeng, H.; Li, G.; Tong, T. Graph Reasoning-Based Emotion Recognition Network. *IEEE Access* **2021**, *9*, 6488–6497. [CrossRef]
36. Huang, D.; Shan, C.; Ardabilian, M.; Wang, Y.; Chen, L. Local Binary Patterns and Its Application to Facial Image Analysis: A Survey. *IEEE Trans. Syst. Man Cybern.* **2011**, *41*, 765–781. [CrossRef]
37. Sahni, T.; Chandak, C.; Chedeti, N.R.; Singh, M. Efficient Twitter sentiment classification using subjective distant supervision. In Proceedings of the 9th International Conference on Communication Systems and Networks (COMSNETS), Bengaluru, India, 4–8 January 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 548–553.
38. Sainath, T.N.; Vinyals, O.; Senior, A.W.; Sak, H. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 4580–4584.
39. Nakov, P.; Ritter, A.; Rosenthal, S.; Sebastiani, F.; Stoyanov, V. SemEval-2016 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT), San Diego, CA, USA, 16–17 June 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 1–18.
40. Tang, D.; Qin, B.; Liu, T. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 1422–1432.
41. Islam, J.; Zhang, Y. Visual Sentiment Analysis for Social Images Using Transfer Learning Approach. In Proceedings of the IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom), Atlanta, GA, USA, 8–10 October 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 124–130.
42. Jou, B.; Chen, T.; Pappas, N.; Redi, M.; Topkara, M.; Chang, S. Visual Affect Around the World: A Large-scale Multilingual Visual Sentiment Ontology. In Proceedings of the 23rd Annual ACM Conference on Multimedia Conference (MM), Brisbane, Australia, 26–30 October 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 159–168.
43. You, Q.; Luo, J.; Jin, H.; Yang, J. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; The AAAI Press: Palo Alto, CA, USA, 2016; pp. 308–314.
44. Mai, L.; Jin, H.; Liu, F. Composition-Preserving Deep Photo Aesthetics Assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 497–506.
45. Ma, S.; Liu, J.; Chen, C. A-Lamp: Adaptive Layout-Aware Multi-patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 722–731.
46. Sheng, K.; Dong, W.; Ma, C.; Mei, X.; Huang, F.; Hu, B. Attention-based Multi-Patch Aggregation for Image Aesthetic Assessment. In Proceedings of the ACM Multimedia Conference on Multimedia Conference (MM), Seoul, Korea, 22–26 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 879–886.
47. Pan, B.; Wang, S.; Jiang, Q. Image Aesthetic Assessment Assisted by Attributes through Adversarial Learning. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; The AAAI Press: Palo Alto, CA, USA, 2019; pp. 679–686.
48. Joshi, D.; Datta, R.; Fedorovskaya, E.A.; Luong, Q.; Wang, J.Z.; Li, J.; Luo, J. Aesthetics and Emotions in Images. *IEEE Signal Process. Mag.* **2011**, *28*, 94–115. [CrossRef]
49. Yu, J.; Cui, C.; Geng, L.; Ma, Y.; Yin, Y. Towards Unified Aesthetics and Emotion Prediction in Images. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE Computer Society: Washington, DC, USA, 2019; pp. 2526–2530.
50. Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; Tian, Q. Deep Modular Co-Attention Networks for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; IEEE Computer Society: Washington, DC, USA, 2019; pp. 6281–6290.
51. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 32, Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 13–23.

52. Yao, L.; Mao, C.; Luo, Y. Graph Convolutional Networks for Text Classification. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, Hawaii, USA, 27 January–1 February 2019; The AAAI Press: Palo Alto, CA, USA, 2019; pp. 7370–7377.
53. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A.F. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) and the 9th International Joint Conference on Natural Language Processing (IJCNLP), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 154–164.
54. Chen, L.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: New York, NY, USA, 2018; pp. 833–851.
55. Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene Parsing through ADE20K Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 5122–5130.
56. DPChallenge. Available online: http://www.dpchallenge.com (accessed on 20 March 2021).
57. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543.
58. Wang, X.; Girshick, R.B.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; IEEE Computer Society: Washington, DC, USA, 2018; pp. 7794–7803.
59. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Advances in Neural Information Processing Systems 29, Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Barcelona, Spain, 5–10 December 2016*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 289–297.
60. Wei, S.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 4724–4732.
61. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human Pose Estimation with Iterative Error Feedback. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 4733–4742.
62. DeviantArt. Available online: https://www.deviantart.com (accessed on 23 March 2021).
63. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 1800–1807.
64. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
65. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 770–778.
66. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; BMVA Press: Guildford, UK, 2016.
67. Hosu, V.; Goldlücke, B.; Saupe, D. Effective Aesthetics Prediction with Multi-Level Spatially Pooled Features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; IEEE Computer Society: Washington, DC, USA, 2019; pp. 9375–9383.
68. Yao, X.; Zhao, S.; Lai, Y.; She, D.; Liang, J.; Yang, J. APSE: Attention-aware Polarity-Sensitive Embedding for Emotion-based Image Retrieval. *IEEE Trans. Multimed.* **2020**. [CrossRef]
69. Zhao, S.; Lin, C.; Xu, P.; Zhao, S.; Guo, Y.; Krishna, R.; Ding, G.; Keutzer, K. CycleEmotionGAN: Emotional Semantic Consistency Preserved CycleGAN for Adapting Image Emotions. In Proceedings of the 33th Association-for-the-Advancement-of-Artificial-Intelligence (AAAI) Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; The AAAI Press: Palo Alto, CA, USA, 2019; pp. 2620–2627.
70. Girshick, R.B. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE Computer Society: Washington, DC, USA, 2015; pp. 1440–1448.
71. Misra, I.; Shrivastava, A.; Gupta, A.; Hebert, M. Cross-Stitch Networks for Multi-task Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington, DC, USA, 2016; pp. 3994–4003.