applied sciences

MDPI

Article

Multitask Learning-Based Affective Prediction for Videos of Films and TV Scenes

Zhibin Su [1,2,3], Shige Lin [3], Luyue Zhang [2,3], Yiming Feng [2,3] and Wei Jiang [1,2,*]

[1] State Key Laboratory of Media Convergence and Communication, Beijing 100024, China; zhibin_su@163.com
[2] Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture and Tourism, Beijing 100024, China; zhangluyue@cuc.edu.cn (L.Z.); fyming8888@163.com (Y.F.)
[3] School of Information and Communication Engineering, Communication University of China, Beijing 100024, China; linshigeaxi@163.com
* Correspondence: jw@cuc.edu.cn

Abstract: Film and TV video scenes contain rich art and design elements such as light and shadow, color, composition, and complex affects. To recognize the fine-grained affects of the art carrier, this paper proposes a multitask affective value prediction model based on an attention mechanism. After comparing the characteristics of different models, a multitask prediction framework based on the improved progressive layered extraction (PLE) architecture (multi-headed attention and factor correlation-based PLE), incorporating a multi-headed self-attention mechanism and correlation analysis of affective factors, is constructed. Both the dynamic and static features of a video are chosen as fusion input, while the regression of fine-grained affects and classification of whether a character exists in a video are designed as different training tasks. Considering the correlation between different affects, we propose a loss function based on association constraints, which effectively solves the problem of training balance within tasks. Experimental results on a self-built video dataset show that the algorithm can give full play to the complementary advantages of different features and improve the accuracy of prediction, which is more suitable for fine-grained affect mining of film and TV scenes.

Keywords: film and TV scenes; multitask learning; fine-grained affects; affective value prediction

check for updates

Citation: Su, Z.; Lin, S.; Zhang, L.; Feng, Y.; Jiang, W. Multitask Learning-Based Affective Prediction for Videos of Films and TV Scenes. *Appl. Sci.* **2024**, *14*, 4391. https://doi.org/10.3390/app14114391

Academic Editor: Andrea Prati

Received: 28 March 2024
Revised: 19 May 2024
Accepted: 20 May 2024
Published: 22 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The emotional responses and inner feelings distinguished by humans are complex and diverse, ranging from simple emotions to deep content cognitive semantics. It is of great value in the imitation of human perception to identify different situations in the application of multimedia products. As the main carrier of human perception, visual affective analysis can better explain the interpretation of visual information by identifying the affects and emotions expressed in the color, composition, facial expressions, and actions in images and videos. A variety of intelligent apps benefit from video affective recognition, such as automatic drives, educational robots, rehabilitation for people with autism, and audiovisual interaction.

In practical applications, video affective prediction for complex scenes is required. However, the most advanced video affect analysis technology shows excellent performance in a single scene, but video affective value prediction for complex scenes is an outstanding problem. Film and TV videos and images, which involve characters, architecture, landscapes, and animals, contain rich emotional information. Therefore, this paper adopts film and TV scenes as the research object for affective value prediction. Most audiovisual affective analysis studies adopt coarse-grained sentiment to obtain a more intuitive sentiment orientation such as binary polarity [1,2], triple classification [3], or basic human emotions [4,5]; however, the impressed affects that the audience can feel from film and

TV scenes are not simple, and thus, the polarity classification of only one identified category cannot describe the feelings of the audience. We investigate a fine-grained affective prediction method to capture the perceived details of affects for film and TV scenes.

Mainstream methods of affective analysis are based on a single-task network architecture [6–10]. Achieving more complex affective prediction requires multiple training models in parallel, which increases the time and complexity of calculation. The affects of the same polarity have similar underlying characteristics and are well-suited to be obtained through simultaneous training. Other elements in the scene, such as composition and the presence of characters, are also important factors affecting the distribution of affects, which can assist in model building. In this context, multitask learning is one of the best solutions for fine-grained affective prediction, which has proved to successfully enhance the performance of individual tasks with the inclusion of other correlated tasks in the training process [11]. Pons et al. [12] employ a CNN as a shared layer, focus on multitask training of image emotion 7 classification and facial action units, define loss functions separately for each task, and find that emotion recognition can benefit from joint learning models. Zhao et al. [13] formalize the emotion distribution prediction task as a shared sparse regression problem and extend it to multitask learning, named multitask shared sparse regression (MTSSR), to explore the latent information between different prediction tasks. J. Shen et al. [14] consider the potential mapping relationship between two emotion models in videos. A framework for multimodal sentiment recognition in videos using multitask learning is proposed, where the two tasks are defined as categorical sentiment classification and dimensional sentiment regression, respectively, and a multitask learning architecture is used to improve the performance of sentiment recognition.

Thus, tasks with small amounts of data available can benefit from being trained simultaneously with other tasks, sharing a common feature representation and transferring knowledge between different domains [15–17]. One of the main difficulties of multitask learning is the joint training process for balancing all tasks. To solve this problem, classical methods split each task with separate loss functions and sum all the losses [12]. However, the different magnitudes of tasks may cause the model to be overly biased for some tasks while ignoring the effects of others; the correlation between tasks is also ignored, resulting in poor overall results for multiple tasks [18,19]. To obtain the correlation among different tasks, a joint loss algorithm based on factor analysis (FA) is proposed.

The contributions of this work include the following:

1. A modified multitask learning model with gating network architecture (multi-headed attention and factor correlation-based progressive layered extraction (MHAF-PLE)) is proposed for fine-grained affective prediction. The additional label of the presence of characters in the picture becomes an auxiliary task to improve the prediction effect of the model.
2. The static information of film and TV scene video keyframes is appended to the video dynamic information to obtain more complete visual information. Using multi-headed self-attention, features with rich information can be given higher weights, which is proved to have a better affective representation ability.
3. A mixed loss function based on factor association constraints is proposed, giving the same weight to sentiments with strong relevance and combining the change in weights and losses.

2. Related Work

2.1. Affective Analysis Model

In addition to a specific human emotional state, an affect can refer to any sensory, physical, mental, or spiritual feeling that can be conveyed and expressed through language. Analyzing and understanding a user's emotional state is a necessary condition for artificial intelligence, emotional computing, and human-computer interaction. Common emotion representation models include Mikel's emotion wheel [12] and Plutchik's wheel of emotions [20]. However, the emotions embedded in an image or video are rich, and

to simply classify them into a certain category is insufficiently precise. The six or eight common categories do not contain enough emotions, while the labels of the cyclic emotion or PAD [21] model cannot clearly show the emotion of the target carrier. To increase the predictive ability for complex scenarios, this paper subdivides affects based on existing sentiment models.

2.2. Visual Emotional Features

Visual affective prediction is classified from the perspective of features and involves methods based on low-level visual features, high-level semantic features, and deep learning features.

Affective prediction methods based on low-end visual features extract low-level visual features from video images. Commonly used bottom-level feature extraction methods for the prediction of emotions in video images include GLCM [22] texture features, SIFT [23] features, and Tamura [24] features. Machajdik et al. [25] combined psychology and art theory in a method that contains color, texture variations, and simple semantic information. Lu et al. [26] investigated the influence of shape features, such as straight lines and curves, on sentiment classification.

Semantic feature-based affective prediction methods translate high-level semantic information into affect features so as to establish a connection between video image sentiment and semantics such as objects and scenes. Borth et al. [27] screened 1200 adjective-noun pairs (ANPs) and constructed a 1200-dimensional feature vector to detect concepts in images and indirectly express image sentiment. Gao et al. [13] used a topic model to extract the emotional features of a movie.

The machine learning-based approach performs affective analysis by extracting visual features of video images and analyzing the differences between visual low-level features and high-level semantics, which effectively reduces the neglect of affective information. The convolutional neural network [12] has subsequently become an important technique in computer vision due to features such as weight sharing and translation invariance and has been widely used for visual emotion feature extraction in recent years. Fayyaz et al. [28]. proposed the use of a convolutional neural network (CNN) to extract spatial features and long short-term memory for feature extraction of time series.

Most emotions are embedded within the images, but factors such as the pace of motion, changes in camera angles, and the narrative structure also influence the emotional response during viewing. In the extraction of video features related to emotions, Al-Saadawi et al. [6] used a CNN to process video data for visual feature extraction, focusing on analyzing facial expressions and body language. In recent years, researchers have begun to apply Transformers to computer vision tasks to improve the quality of feature representation by capturing long-distance dependencies between frames. Yang et al. [3] use a transformer encoder for video vision feature extraction to capture the dependencies between frames.

Traditional affect analysis work considers either video dynamic information or static information of video keyframes and does not fully exploit the joint visual expression of video. We use video-image fusion features to improve video affective value prediction, obtain more complete visual information, and enhance the accuracy of fine-grained affective prediction.

2.3. Multitask Learning

Multitask learning is a migration learning method that aims to improve the performance of a single task by simultaneously learning multiple related tasks [29]. The basic idea is to maximize information sharing and migration ability between tasks by introducing a shared layer in the network, where added tasks change the weights and update directions of parameters, thus improving the generalization performance of the classifier [16,30], especially in the case of small sample sizes [17,31,32]. An important feature of multitask learning is its ability to capture correlations between tasks by learning shared representations, thus improving overall system performance.

Figure 1a shows the structure of hard parameter, hard parameter sharing is the most basic and commonly used MTL structure, but because parameters are shared directly among tasks [33], they may suffer from negative migration due to task conflicts. MoE [34] addressed these by sharing some experts at the bottom layer and combining expert groups through gated networks. Figure 1b shows the structure of MMoE, MMoE [35] extends MoE [34] using different gates for each task so as to obtain different fusion weights in MTL. However, prior knowledge about the network design is important, and MMoE [35] has difficulty discovering the path of fusion in practice. To represent the joint optimization of learning and tasks, especially in a non-separable joint approach. Figure 1c shows the structure of PLE, PLE [36] uses a progressive separation approach on the general framework of joint learning and tasks, which adds an exclusive expert network to focus more on the exclusive mode of learning tasks while retaining the shared expert network. However, the PLE model does not balance well the loss functions of multiple tasks nor exploit well the unique features within the images and videos themselves, and it still suffers from negative migration and the seesaw phenomenon when the number of tasks is high. Therefore, we propose a loss function based on the factor association constraint by combining the representation learning of feature weights with the attention mechanism in the PLE [36] model framework. An improved multitask fine-grained affective prediction model is constructed to solve the feature representation problem and the seesaw phenomenon and to balance the learning of multiple tasks.

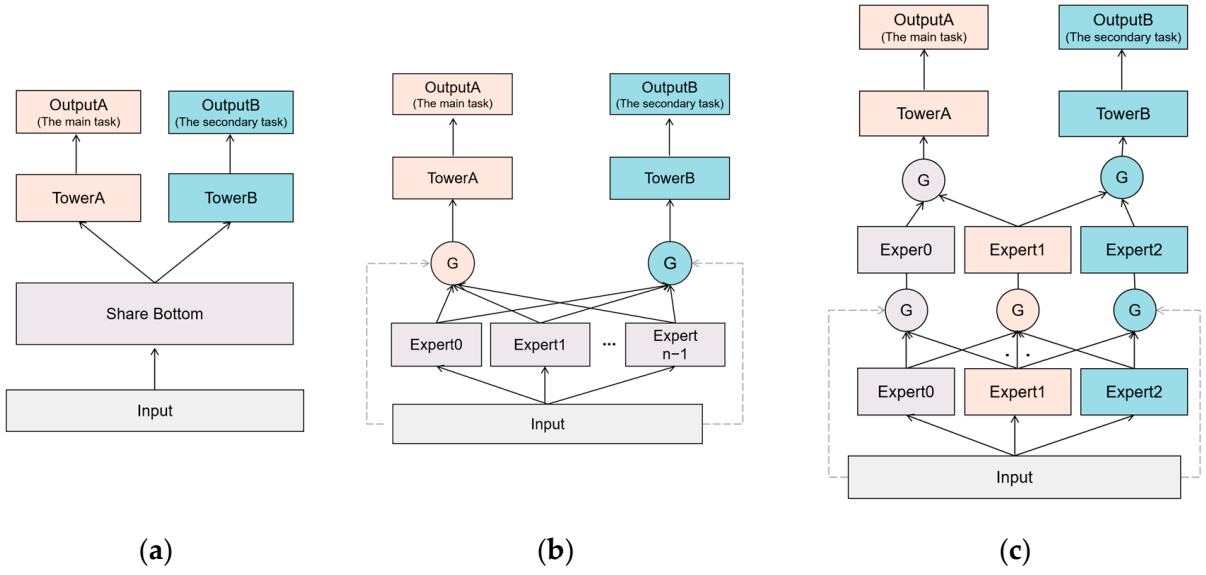


Figure 1. Most commonly used MTL structures. (**a**) Description of the structure of Shared-Bottom; (**b**) Description of the structure of MMoE; (**c**) Description of the structure of PLE.

3. Method

3.1. Proposed Framework

The suggested MTL is a strategy for machine learning in which n learning tasks are simultaneously executed using commonalities and differences across tasks [37]. The existing models are computation-intensive and take much time to separately process different tasks. We subdivided human emotions into 16 emotion adjectives and performed 16 emotion prediction tasks, taking the emotion prediction task as the main task. The 16 emotion adjectives involved warm, magnificent, depressed, happy, relaxed, anxious, dreamy, hopeful, sentimental, sunny, romantic, oppressive, fresh, cozy, disappointed, and lonely.

Figure 2 shows the structure of the proposed MHAF-PLE model, including stages of feature extraction, feature refinement, and joint training. The feature-sharing layer model is based on the PLE framework. The feature extractor that performs best in the main task

is first selected for feature extraction via a feature refinement layer with multi-headed attention as the core in order to learn richer and more comprehensive transferable visual representations. The affects of the same sentiment dimension are given the same weight in combination with FA to learn at a similar rate and avoid neglecting the influence of other tasks due to the overweighting of one affect task. We describe each module in the following text.

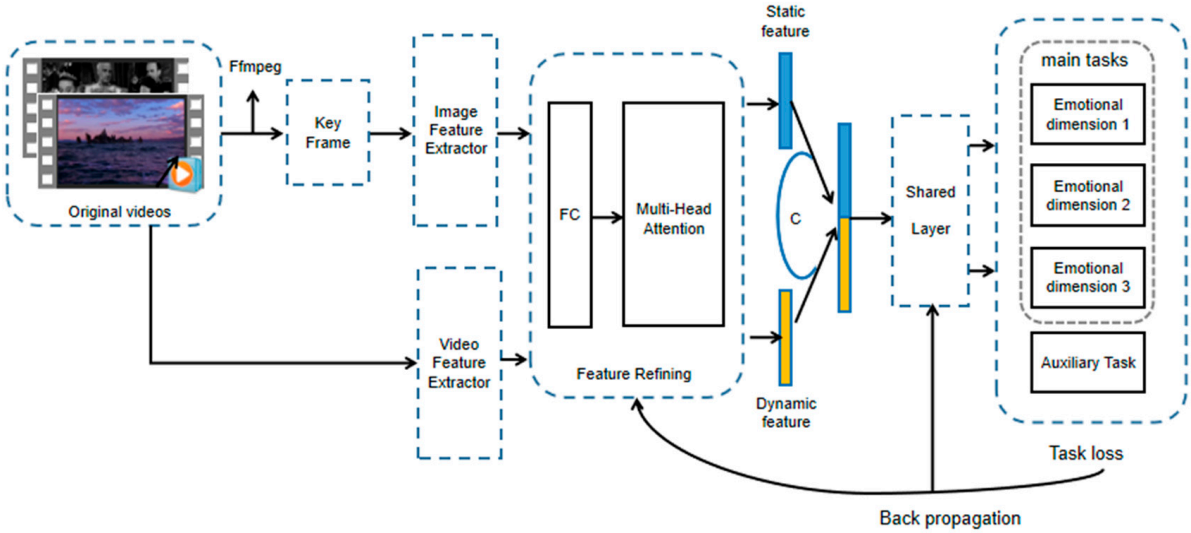


Figure 2. Proposed framework: MHAF-PLE.

3.2. Feature Extraction

3.2.1. Video Dynamic Feature Extraction

In the video dynamic feature extraction stage, because the extraction of emotional features is performed from video scenes that change over time, it is only necessary to extract image frames at certain time intervals, and the spatiotemporal emotional information in these frames can be analyzed and extracted without processing the entire video, improving computational efficiency.

We use ffmpeg to extract keyframes with video context information [38]. Here, the keyframes represent the first and last frames of each full video sequence. After the first step, the extracted frames are fed into the pre-trained dynamic feature extractor, including the TimeSformer model, the MviT model, and the widely used C3D model. The TimeSformer input is 8 video frames of size 224 × 224 extracted from the video; output is 400-dimensional video frame features per frame. It utilizes a pre-trained model trained on the kinetics-600 dataset, with attention mechanisms applied in both spatial and temporal dimensions. The backbone network is a Vision Transformer (ViT) based video module with 12 layers. MviT outputs 1000-dimensional features per frame, and the C3D model outputs 496-dimensional features. The mean absolute errors (MAEs) of different video dynamic feature extractors for the main task are shown in Table 1, with the best performance marked in bold. The TimeSformer model uses a spatiotemporal separation of attention mechanisms to extend the image space into spatiotemporal 3D space. We first consider attention between patches in the same spatial location in the temporal dimension and then consider attention in the same intra-frame space. This allows us to extract the affective information of video frames at specific time intervals containing contextual information without the need to extract features from the whole video, thereby improving computational efficiency. Therefore, the TimeSformer model, which has the best performance on the main task, was selected after experimental comparison.

Table 1. Prediction results based on different video dynamic feature selection methods.

Model	MAE of Affective Prediciton
MviT	0.1898
C3D	0.2484
TimeSformer	**0.1637**

3.2.2. Video Static Feature Extraction

The first frame images of the video are selected, and each image is scaled to a fixed size of 224 × 224 and then directly fed into the pre-trained image feature extractors Vgg16 [19], GoogLeNet [39], ViT [40], and CLIP [41] with the network framework of keras. We selected the best-performing 512-dimensional output among the 512, 1024, 2048, and 4096-dimensional features of the pre-trained VGG16 output, the 1000-dimensional features of the pre-trained GoogleNet output, and the 1000-dimensional features of the pre-trained ViT output. CLIP output is 512-dimensional features, the selected pre-trained model for ClIP is a pre-trained model on a dataset of 400 million images, and the backbone network is the Vision Transformer (ViT) with a total of 16 layers. The experiments all use the same experimental environment and experimental parameters; the optimizer is adam, the learning rate is 0.0002, the dropout is 0.2, the batch size is 50, and the number of epochs is 200. In addition, in order to improve the persuasiveness of the experimental results, this paper adopts the 5-fold cross-validation method.

The mean absolute errors (MAEs) of different feature extractors for the main task are shown in Table 2, with the best performance marked in bold. Therefore, the CLIP model, which has the best performance on the main task, is selected after experimental comparison.

Table 2. Prediction results based on different video static feature selection methods.

Model	MAE of Affective Prediction
Vgg16	0.1057
GoogLeNet	0.0950
ViT	0.1432
CLIP	**0.0885**

3.3. Feature Refining

Because direct feature concatenation alone cannot effectively mine the unique features within the videos and images themselves, using a single gated network may result in the network overly emphasizing similar parts and failing to mine the unique information within the features. Therefore, the work of fine-grained affective prediction of video in complex scenes must address the heterogeneity of video and image features and mine the unique feature information within the videos and images themselves. Based on this, we introduce a feature refinement module to give higher weights to the information-rich features, as shown in Figure 3.

The feature refinement layer accepts the output of the feature extraction layer, which has video features containing dynamic features and image features containing static features; the features have different semantics and belong to different spaces. We use the feature refinement module to represent video and image features and mine their unique feature information. A multi-headed attention mechanism uses attention weight calculation to control feature flow and integration. Attention mechanism-based models have been shown to universally enhance the effectiveness of features [42–47]. We use a multi-headed mechanism [48], which realizes learning features from different dimensional semantic representation subspaces, improving fine-grained affective prediction accuracy.

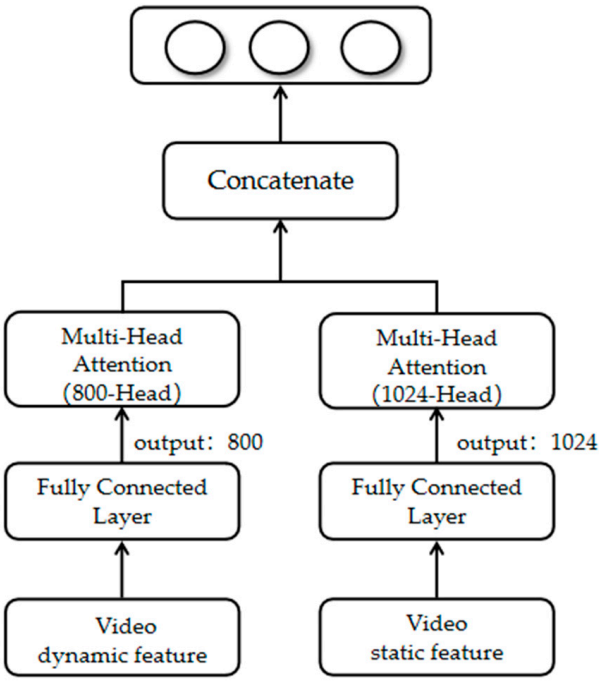


Figure 3. Proposed framework: Feature Refining.

Video and image features are input separately to the fully connected layer, whose output is subjected to multi-headed self-attention, which uses multiple attention heads to compute the similarity between query and index vectors in different subspaces and extract richer contextual information. Multi-headed attention can learn richer feature combinations using different dimensional semantic feature subspaces, preventing overfitting by integrating the results of multiple mutually independent attention computations and thus improving fine-grained affective prediction results.

3.4. Loss Function

3.4.1. Factor Analysis

The process of FA can be divided into two stages: factor extraction and factor rotation. In the first stage, the original variables are transformed into a set of common factors that explain the common variance of the original variables, such as by the maximum likelihood, minimum residual, or principal component method. In the second stage, the extracted factors are rotated so that each more clearly represents a set of original variables.

Standard FA analysis [49] aims to extract the potential common variables from multiple parameters, while this paper designs a loss function based on the FA method. Specifically, the magnitude of loss associated with larger load values for the same common factors should be as close as possible to each other. Therefore, affects with stronger correlations are given the same weight, which ensures that they learn at a similar speed in the training process. Each affect dimension can be represented by a linear combination of a few potential factors plus a special factor for each affect dimension.

The maximum variance orthogonal rotation method in principal component analysis was chosen to test for changes in the number of factors. By analyzing the variance explained by principal components, the variables that contribute to the total variance can be identified so as to better understand the characteristics and structure of the data. Components with eigenvalues greater than 1 are selected as principal components, and appropriate factors are selected based on the cumulative variance ratio. The number of factors selected should generally include more than 60% of the total variance. The principal component results of the FA of affect using the truth values of affect words are shown in Table 3.

Table 3. Results of factor analysis of 16 affective adjectives.

Component	Factor 1	Factor 2	Factor 3
Warm	**0.913**	−0.298	0.100
Hopeful	**0.798**	−0.352	0.275
Happy	**0.872**	−0.335	0.005
Romantic	**0.65**	−0.241	0.48
Relaxed	**0.839**	−0.449	0.126
Fresh	**0.711**	−0.357	0.359
Cozy	**0.908**	−0.299	0.015
Sunny	**0.872**	−0.326	0.151
Magnificent	−0.054	−0.239	**0.849**
Dreamy	0.432	−0.235	**0.791**
Lonely	−0.397	**0.795**	−0.085
Sentimental	−0.269	**0.897**	−0.235
Disappointed	−0.259	**0.904**	−0.232
Depressed	−0.349	**0.893**	−0.209
Oppressive	−0.511	**0.739**	−0.21
Anxious	−0.427	**0.827**	−0.247
Indicator			
Eigenvalue	10.863	1.826	1.208
Variance (%)	67.897	11.414	7.551
Cumulative (%)	67.863	79.311	86.862

Table 3 shows that the eigenvalues of the three extracted factors are greater than 1, and the total cumulative variance ratio is 86.7%, indicating that the FA results using the three factors are credible. The corresponding maximum absolute value of the factor load is marked in bold in Table 3. The first common factor shows larger loadings on warm, hopeful, happy, romantic, relaxed, fresh, cozy, and sunny emotions than on other emotions, indicating that these eight emotions have a strong correlation and are considered positive-class emotions. The second common factor shows larger loadings on loneliness, sadness, loss, disappointment, depression, and apprehension than on other emotions, indicating that these six emotions have a strong correlation and are regarded as negative-class emotions. The third common factor shows larger loadings on magnificent and dreamy emotions than on other emotions, indicating that these two emotions are strongly correlated and can be considered scene-class emotions.

Because the 16 emotions can be divided into three dimensions, the same loss weight is assigned to emotions within the same dimension for multitask learning weight optimization so as to improve the accuracy of fine-grained emotion prediction for movie and TV scenes.

3.4.2. Joint Loss Function

We adopt a multitask modeling approach, with joint primary and secondary tasks for fine-grained affective prediction. In the main task, the loss function is based on the extraction of affective dimensions from the FA, and tasks of the same category are learned at the same speed. Hence, the tasks of each category have the same degree of importance and receive the same attention, balancing the loss and avoiding seesaw phenomena [50–52]. MAE is used for the affective regression task and categorical cross-entropy loss for the classification task. The joint loss function to be minimized is

$$L = W_{pos}\sum_{1}^{n_p} loss_{pos} + W_{sce}\sum_{1}^{n_s} loss_{sce} + W_{neg}\sum_{1}^{n_n} loss_{neg} + \partial loss_{pp}, \quad (1)$$

where $Loss_{pos}$, $Loss_{sce}$, $Loss_{neg}$, and $Loss_{pp}$ are the respective losses incurred in the positive emotions, scene emotions, negative emotions, and classification tasks.

When training a multitask network, it is difficult to balance the training of these tasks, and the change rate of loss can reflect the difficulty of learning a task, so the average value of the loss change rate of different emotion tasks is considered. Considering the average value of the loss change rate of the same dimension of emotion after a certain number of

rounds of training, the weight of corresponding tasks can be changed, which connects different categories of emotions and reduces the dominance of a certain category of emotion task, which could result in an inability to converge. The ratio of the mean values of loss change rates for different dimensions after a certain number of training rounds is

$$Loss_{pos} : Loss_{sce} : Loss_{neg} = K_{pos} : K_{sce} : K_{neg}, \quad (2)$$

The respective weights of positive emotion, scene emotion, and negative emotion for a proposed scheme a are defined as follows:

$$W_{pos} = aK_{pos}, \quad (3)$$

$$W_{sce} = aK_{sce}, \quad (4)$$

$$W_{neg} = aK_{neg}, \quad (5)$$

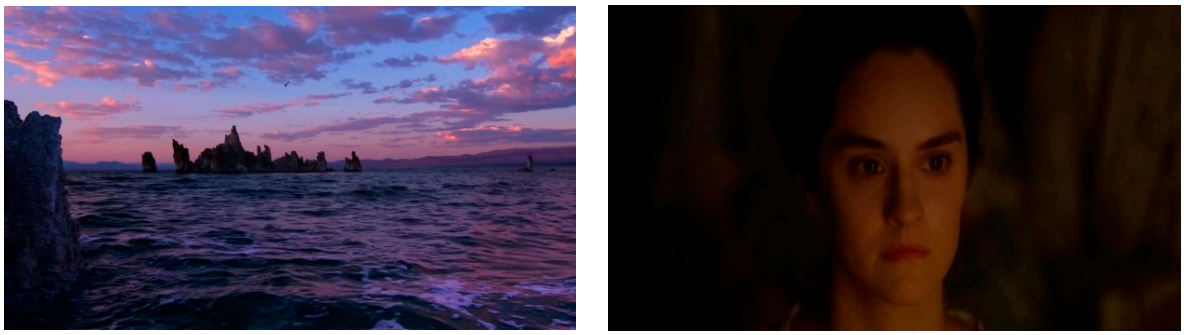
where the multiplicative factors K_{pos}, K_{sce}, and K_{neg} regulate the magnitudes of the weights.

4. Experimental Results

4.1. Dataset

Public video affective datasets are mainly multimodal. MOSI is a widely-used dataset comprising 2199 discourse video clips from 93 YouTube videos, containing six aspects of emotion labeled as happy, sad, angry, fearful, disgusted, and surprised, with scores for each label ranging from −3 (strongly negative) to +3 (strongly positive). The MOSEI dataset is the larger version of MOSI, containing 22856 annotated video clips on 250 different topics. The labels are the same as MOSI. CH-SIMS is a Chinese unimodal and multimodal affective analysis dataset that contains 2281 refined video clips. These video clips are collected from different movies, TV series, and variety shows. Each sample has one multimodal label and three unimodal labels, and the affective scores are categorized into three categories: −1 (negative), 0 (neutral), and 1 (positive) [3]. However, the unimodal labels only contain positive, neutral, and negative categories and cannot be well adapted to the fine-grained affect analysis task of this paper, so we choose to self-construct the film and television scene dataset containing fine-grained affective labels, which can better segment and deeply analyze the affect of the art carriers.

We self-constructed a movie and TV scene video dataset whose samples consisted of long-shot clips extracted, such as from existing movie and TV scene videos, architecture, landscape, and animals, involving a wide range of scenes, characters, and landscapes, with half the labeling balanced. The sample video intercepted the first frame image, as shown in Figure 4.



(**a**) (**b**)

Figure 4. Example images of video dataset. (**a**) Scene without character; (**b**) Scene with character.

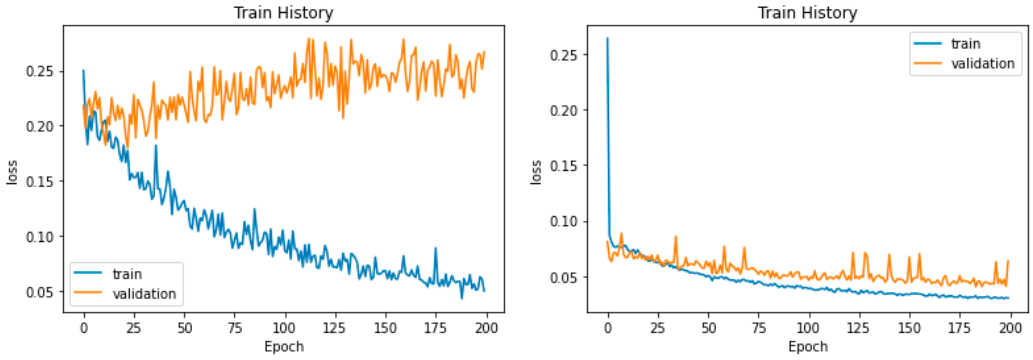
Table 4 shows the 16 affective adjectives [53] for the fine-grained prediction in this paper. On this basis, a subjective evaluation experiment was conducted, and the score of

each adjective was the average of 30 subjects after removing outliers. The corresponding emotion label for each sample was the real value of $y \in [0.0, 5.0]$, which is the emotion score.

Table 4. Set of emotional adjectives.

Warm	Magnificent	Depressed	Happy
Relaxed	Anxious	Dreamy	Hopeful
Sentimental	Sunny	Romantic	Oppressive
Fresh	Cozy	Disappointed	Lonely

Four ways of sampling were used to obtain 1600 video clips of movie and TV videos with a length of 10 s. After sampling the frames, each image was compressed to a fixed size of 224 × 224 pixels. Figure 5 shows the training–validation loss plot with or without data enhancement, which indicates that the proposed data enhancement method effectively prevents overfitting.



(**a**) (**b**)

Figure 5. Loss fitting curve. (**a**) Raw data; (**b**) Enhanced data.

4.2. Implementation

All experiments were performed in Python, PyTorch was applied for video dynamic feature extraction, and Keras was applied for model implementation. During training, the ADAM optimizer was used, with a learning rate of 0.0002. The training was performed 200 times with a batch size of 200. The loss function weight setting was designed to compute the average loss for different affective dimensions over 50 iterations.

In the experiment, affective value prediction was a regression task trained and evaluated with MAE loss, and in auxiliary tasks, classification was evaluated with a cross-entropy loss function and AUC. In the movie and TV scene video dataset, the training and testing sets were in a 4:1 ratio. To better represent the videos, dynamic and static features of the first and last frames were fused separately.

4.3. Results

4.3.1. Algorithm Prediction Results Based on Different Feature Selection Methods

To better select the video characterization method, experiments were conducted on the fine-grained affect regression task and the character landscape binary classification task group. We fused the intermediate frame features for experimental comparisons by concatenating the first and last features. Image feature extraction uniformly used the pre-trained CLIP model, while video feature extraction uniformly used the pre-trained TimeSformer model. The experiments were conducted on a self-constructed dataset of film and television scene videos.

The experimental results are shown in Table 5, with the best scores marked in bold. To observe the effect of multi-frame fusion on training, we output the update process of loss in different cases, as shown in Figure 5.

Table 5. Prediction results based on different feature selection methods.

			TimeSformer + CLIP	
	First Frame		Frame	
Component	**TimeSformer**	**TimeSformer + Clip**	**First + Last**	**First + Middle + Last**
Warm	0.0522	0.0248	0.0270	0.0254
Hopeful	0.1754	0.0713	0.0606	0.0621
Happy	0.1936	0.0715	0.0548	0.0565
Romantic	0.1613	0.0545	0.0507	0.0529
Relaxed	0.2080	0.0793	0.0571	0.0597
Fresh	0.1646	0.0577	0.0581	0.0593
Cozy	0.1877	0.0581	0.0514	0.0527
Sunny	0.1901	0.0694	0.0589	0.0602
Magnificent	0.1331	0.0508	0.0472	0.0481
Dreamy	0.1613	0.0526	0.0472	0.0513
Lonely	0.1887	0.0707	0.0643	0.0661
Sentimental	0.1556	0.0661	0.0643	0.0650
Disappointed	0.1536	0.0676	0.0546	0.0551
Depressed	0.1518	0.0670	0.0519	0.0523
Oppressive	0.1686	0.0737	0.0506	0.0643
Anxious	0.1731	0.0651	0.0580	0.0612
Average (MAE)	0.1637	0.0694	**0.0542**	0.0558
Classification (%)	76%	97%	**98%**	97%

There are three findings from Table 5. Relative to the single video dynamic feature, the mean value of the MAE for the main emotion prediction task using video image fusion features was reduced by 0.1012, and the MAE values for each emotion task were also reduced, with coziness showing the largest reduction of 0.1296 and warmth the smallest reduction of 0.0274. The accuracy of the auxiliary task was improved by 20.75%. Therefore, video image fusion features were used for the subsequent study. Relative to the single-frame feature, the multi-frame fusion feature impaired three affective tasks but improved 13 affective tasks, as well as the auxiliary task; the mean value of the MAE for the affective regression task was reduced by 0.0353, and the accuracy of the auxiliary task was improved by 1%. Compared with the fusion feature of adding intermediate frames, the first and last frame fusion features impaired the loss regression task but improved the rest of the tasks, and the mean value of the MAE for the emotion regression task was reduced by 0.0016. The best performance was achieved using the first and last frame fusion features. This shows that multi-frame fusion features can form a clearer, more complete, and more accurate description of the video, expand spatiotemporal information, reduce uncertainty, increase reliability, and improve the accuracy of fine-grained affective prediction. Fusing the first and last frame features instead of all frames can reduce the uncertainty and redundancy of the output based on the maximum merging of relevant information.

4.3.2. Algorithm Prediction Results Based on Different Task Selection Methods

To verify the effectiveness of the proposed multitask learning task settings and the applicability of multiple tasks on the proposed model, different emotion tasks were selected for comparison, along with the introduction of other auxiliary tasks. We selected the positive emotion "warmth" as the central emotion, which performed well and added main tasks for predicting positive or negative emotions based on it. Considering the balance between the number of positive and negative emotion tasks, an equal proportion of emotion polarity tasks were added. For the selection of auxiliary tasks, we considered commonly used emotion polarity classification tasks. Because characters and scenery are the main subjects in film and television scene videos, we considered the classification of whether characters are present. The selected tasks are shown in Table 6. Video feature extraction uniformly employs the pre-trained TimeSformer model, and features are uniformly selected as fused features of image and video. Within the framework of multitask learning, we

adopt an expert-gating sharing mechanism to share the underlying feature representation. The experiments are conducted on a self-constructed dataset of film and television scene videos, the corresponding prediction results are shown in Table 7.

Table 6. Selected tasks with corresponding numbers.

Number	Task
1	Positive emotion (warm) predictions + character classification + emotional polarity classification
2	Positive emotion (warm) prediction + negative emotion (disappointed) prediction + character classification
3	Positive emotion (warm and magnificent) prediction + negative emotion prediction (disappointed and depressed) + character classification
4	Positive emotion (warm, magnificent, and sunny) prediction + negative emotion prediction (disappointed and depressed) + character classification

Table 7. Prediction results based on different task selection methods.

	Task Number				
Predict Result	**1**	**2**	**3**	**4**	**Proposed Task**
Warm	0.0370	0.0270	0.0258	0.2500	**0.0248**
Sunny	/	/	0.0781	0.0725	**0.0694**
Magnificent	/	/	0.0600	0.0536	**0.0508**
Disappointed	/	0.0892	0.0754	0.0720	**0.0676**
Depressed	/	/	/	0.7567	**0.6700**
Character Classification	69%	71%	71%	75%	**97%**
Emotional polarity classification	65%	/	/	/	/

As can be seen in Table 7, with the increase in the number of tasks, the MAE for each emotion task decreases, and the prediction effect is significantly improved; similarly, the accuracy of the binary categorization of character landscapes increases with the increase in the number of emotion tasks. This shows that the proposed model is suitable for fine-grained emotion prediction tasks with a large number of tasks and multitask learning of multiple emotion tasks, sharing more information between emotion tasks and improving the overall effect of the model. As indicated in bold in Table 7, when the number of tasks is 17, the MAE for each emotion task decreases, and the accuracy of the binary categorization of character landscapes reaches 97%. Therefore, the proposed model was based on 17 tasks for carrying out research.

4.3.3. Algorithm Prediction Results Based on Different Loss Functions

We designed experiments to verify the effectiveness of the joint loss function combined with FA and the affective weight setting combined with loss means for fine-grained affective prediction. The proposed joint function was compared with different loss weight cases, such as the consistency of all affect loss weights and the comparison of loss weight outlines, with results as shown in Table 8.

The following three findings can be obtained from the comparison of the experimental results in Table 8. First, in the self-constructed film and TV scene dataset, using the joint loss function that combines FA achieves better performance, reducing the MAE by 0.0026 compared with the setting with consistent weights for all affects. Second, using the weight setting that combines the loss means for each category of affects after a certain number of rounds of training is better than the weight setting that combines only FA but not the loss means after a certain number of rounds of training, reducing the MAE by 0.0017. Third, using a large scaling factor, i.e., a large weight, reduces the MAE by 0.0025 compared with a small weight, and the MAE for each affect is mostly reduced. The best performance,

highlighted in bold in Table 8, is achieved when using the combination of three types of loss together. Therefore, we choose weights with large weight outlines for further discussion.

Table 8. Prediction results based on different loss functions.

Emotion Task	Weight Consistency	FA	FA + Loss Weight	FA + Loss Weight + Amplification Factor
Warm	0.0240	0.0210	0.0210	0.0185
Hopeful	0.0545	0.0488	0.0455	0.0465
Happy	0.0518	0.0452	0.0447	0.0394
Romantic	0.0489	0.0482	0.0488	0.0504
Relaxed	0.0586	0.0508	0.0475	0.0465
Fresh	0.0519	0.0521	0.0510	0.0501
Cozy	0.0447	0.0567	0.0587	0.0444
Sunny	0.0525	0.0545	0.0538	0.0523
Magnificent	0.0505	0.0585	0.0529	0.0498
Dreamy	0.0541	0.0507	0.0472	0.0456
Lonely	0.0479	0.0430	0.0427	0.0433
Sentimental	0.0452	0.0369	0.0343	0.0328
Disappointed	0.0542	0.0423	0.0419	0.0418
Depressed	0.0433	0.0429	0.0381	0.0352
Oppressive	0.0464	0.0614	0.0589	0.0586
Anxious	0.0686	0.0469	0.0451	0.0371
Average	0.0500	0.0474	0.0457	**0.0432**

Within a specified range, an increase in the loss weight correlates with decreased MAE for each task. This relationship is evident upon examining the last two columns in Table 8. Specifically, when the weight a is set to 10, the trend of variation in the predictive results for the primary task is delineated in Table 9. The amplification factor k is defined within a range from 1 to 9. As the k progressively increases, the overall MAE reaches its minimum value at k = 5, which is in bold in Table 9. By employing magnification adjustments, it becomes feasible to pinpoint the optimal equilibrium within the primary task of affective prediction and between the primary and auxiliary tasks. However, an excessively large value of k can disrupt this balance, resulting in diminished performance.

Table 9. Prediction results based on different settings of k.

k	MAE of Affective Prediciton
1	0.0432
3	0.0389
4	0.0388
5	**0.0361**
6	0.0387
7	0.0403
8	0.0412
9	0.0432

4.3.4. Comparison with Other Algorithms

Table 10 lists the results of the different comparison methods on the self-constructed dataset. We selected the classical multitask learning framework Share-Bottom, the classical gated network PLE model, the approach of Pons et al. [12], which uses a CNN as a shared layer for sentiment analysis, and the MMTrans-MT network proposed by Jinrui Shen et al. [14], which uses 4 Transformer encoders as a shared layer for sentiment analysis. To compare with the MHAF-PLE model proposed in our paper, the experiments are conducted in the same experimental environment, and the same features are used for comparison. The results show that MHAF-PLE has the best performance, with improvements in the average absolute error for both fine-grained sentiment prediction for the primary task and

character categorization for the secondary task. The MAE of MHAF-PLE is reduced by 0.1800, 0.1362, 0.1534, and 0.1538 compared to Share-Bottom, PLE, RF [12], and MMTrans-MT [14] models, respectively; in addition, the classification accuracy for the secondary task is also improved by 17%, 18%, and 43%, 39% compared to Share-Bottom, PLE, RF [12], and MMTrans-MT [14] models, respectively. The MAE for each sentiment is also reduced, which suggests that compared to Share-Bottom and PLE models, RF [12], and MMTrans-MT [14], the proposed method performs well in fine-grained affective analysis.

Table 10. Prediction results based on different models.

Task	Share-Bottom	PLE	RF. [12]	MMTrans-MT [14]	MHAF-PLE (Ours)
Warm	0.2498	0.1761	0.1619	0.2069	0.0185
Hopeful	0.2329	0.1870	0.1888	0.2027	0.0465
Happy	0.2787	0.1880	0.2252	0.2243	0.0394
Romantic	0.2217	0.1724	0.1868	0.1870	0.0504
Relaxed	0.2695	0.2315	0.1856	0.2295	0.0465
Fresh	0.2555	0.1733	0.2289	0.1945	0.0501
Cozy	0.2603	0.1810	0.1936	0.2123	0.0444
Sunny	0.2503	0.1891	0.1830	0.2163	0.0523
Magnificent	0.1341	0.1025	0.1697	0.1601	0.0498
Dreamy	0.1836	0.1510	0.2062	0.1852	0.0456
Lonely	0.2636	0.2178	0.2126	0.1907	0.0433
Sentimental	0.1784	0.1578	0.2027	0.1831	0.0328
Disappointed	0.1780	0.1637	0.1928	0.1698	0.0418
Depressed	0.1900	0.1757	0.1905	0.1933	0.0352
Oppressive	0.2159	0.2245	0.2147	0.1905	0.0586
Anxious	0.2101	0.1801	0.2038	0.2047	0.0371
Average	0.2232	0.1794	0.1966	0.1970	0.0432
Character classification	80%	79%	55%	59%	98%

The MHAF-PLE model integrates the attention mechanism and gives higher weight to information rich in both video and images, enabling better mining of emotional feature representation. The constraint design based on factor association can adjust the learning speed and proportion through the loss function according to the characteristics of fine-grained affects, improving the accuracy of prediction.

5. Conclusions and Future Work

Considering the fine-grained affective prediction problem of film and TV scene videos, we proposed a multitask prediction model, MHAF-PLE. The algorithm combines dynamic video information with keyframe image information as the input, enhancing the PLE-based gating network with a multi-headed attention mechanism as the primary algorithmic framework. A joint loss function was designed based on factor correlation, ensuring the same learning speed for special affects with strong correlation. The experimental results showed that detecting the presence of characters in the video is an important auxiliary task for affective prediction. MHAF-PLE achieves improved accuracy by adding this task.

The affective attributes of film and TV video are reflected at a multi-dimensional level, especially the high-dimensional characteristics of cognition and perception. However, they are not clearly determined. In the future, we will continue to explore the hidden visual information representation of time and space of pictures from the perspective of designed artworks. In addition, it would be valuable to find the correlation between objective parameters and perceptual attributes through an artificial model for both prediction and artistic appreciation.

Author Contributions: Conceptualization, Z.S.; methodology, Z.S., S.L. and W.J.; software, S.L. and Y.F.; validation, S.L. and L.Z.; formal analysis, L.Z. and W.J.; writing—original draft preparation, Z.S. and S.L.; writing—review and editing, Z.S. and L.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (Grant No. 62276240) and the Fundamental Research Funds for the Central Universities (Grant No. CUC23ZDTJ010).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request. The data are not publicly available due to ongoing research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zhao, S.; Yao, X.; Yang, J.; Jia, G.; Ding, G.; Chua, T.S.; Schuller, B.W.; Keutzer, K. Affective image content analysis: Two decades review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6729–6751. [CrossRef] [PubMed]
2. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [CrossRef]
3. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 7617–7630. [CrossRef]
4. Liu, C.; Zhao, S.; Luo, Y.; Liu, G. TransIEA: Transformer-Baseartd Image Emotion Analysis. In Proceedings of the 2022 7th International Conference on Computer and Communication Systems (ICCCS), Wuhan, China, 22–25 April 2022; pp. 310–313. [CrossRef]
5. Yan, M.; Lou, X.; Chan, C.A.; Wang, Y.; Jiang, W. A Semantic and Emotion-based Dual Latent Variable Generation Model for a Dialogue System. *CAAI Trans. Intell. Technol.* **2023**, *8*, 319–330. [CrossRef]
6. Al-Saadawi, H.F.T.; Das, R. TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network. *Appl. Sci.* **2024**, *14*, 2252. [CrossRef]
7. Karuthakannan, U.K.D.; Velusamy, G. TGSL-Dependent Feature Selection for Boosting the Visual Sentiment Classification. *Symmetry* **2021**, *13*, 1464. [CrossRef]
8. Chaudhari, A.; Bhatt, C.; Krishna, A.; Mazzeo, P.L. ViTFER: Facial Emotion Recognition with Vision Transformers. *Appl. Syst. Innov.* **2022**, *5*, 80. [CrossRef]
9. Zisad, S.N.; Chowdhury, E.; Hossain, M.S.; Islam, R.U.; Andersson, K. An Integrated Deep Learning and Belief Rule-Based Expert System for Visual Sentiment Analysis under Uncertainty. *Algorithms* **2021**, *14*, 213. [CrossRef]
10. Yu, Y.; Lin, H.; Meng, J.; Zhao, Z. Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks. *Algorithms* **2016**, *9*, 41. [CrossRef]
11. Vandenhende, S.; Georgoulis, S.; Proesmans, M.; Dai, D.; Gool, L.V. Revisiting Multi-Task Learning in the Deep Learning Era. *arXiv* **2020**, arXiv:2004.13379v1.
12. Pons, G.; Masip, D. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv* **2018**, arXiv:1802.06664.
13. Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; Ding, G. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE Trans. Multimed.* **2016**, *19*, 632–645. [CrossRef]
14. Shen, J.; Zheng, J.; Wang, X. MMTrans-MT: A Framework for Multimodal Emotion Recognition Using Multitask Learning. In Proceedings of the 13th International Conference on Advanced Computational Intelligence (ICACI), Wanzhou, China, 14–16 May 2021; pp. 52–59. [CrossRef]
15. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef] [PubMed]
16. Maninis, K.K.; Radosavovic, I.; Kokkinos, I. Attentive single-tasking of multiple tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 1851–1860. [CrossRef]
17. Guo, J.; Li, W.; Guan, J.; Gao, H.; Liu, B.; Gong, L. SIM: An improved few-shot image classification model with multi-task learning. *J. Electron. Imaging* **2022**, *31*, 033044. [CrossRef]
18. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? *ICML* **2021**, *2*, 4.
19. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

20. Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of Emotion*; Plutchik, R., Kellerman, H., Eds.; Academic Press: Cambridge, MA, USA, 1980; pp. 3–33. [CrossRef]
21. Mehrabian, A.; O'Reilly, E. Analysis of personality measures in terms of basic dimensions of temperament. *J. Pers. Soc. Psychol.* **1980**, *38*, 492–503. [CrossRef]
22. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man. Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
23. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
24. Losson, O.; Porebski, A.; Vandenbroucke, N.; Macaire, L. Color texture analysis using CFA chromatic co-occurrence matrices. *Comput. Vis. Image Underst.* **2013**, *117*, 747–763. [CrossRef]
25. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM international conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 83–92. [CrossRef]
26. Lu, X.; Suryanarayan, P.; Adams, R.B., Jr.; Li, J.; Newman, M.G.; Wang, J.Z. On shape and the computability of emotions. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 229–238. [CrossRef]
27. Borth, D.; Chen, T.; Ji, R.; Chang, S.F. Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 459–460. [CrossRef]
28. Fayyaz, M.; Saffar, M.H.; Sabokrou, M.; Fathy, M.; Klette, R.; Huang, F. STFCN: Spatio-temporal FCN for semantic video segmentation. *arXiv* **2016**, arXiv:1608.05971.
29. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
30. Bi, Y.; Xue, B.; Zhang, M. Multitask feature learning as multiobjective optimization: A new genetic programming approach to image classification. *IEEE Trans. Cybern.* **2022**, *53*, 3007–3020. [CrossRef] [PubMed]
31. Hou, Y.; Lai, Y.; Chen, C.; Che, W.; Liu, T. Learning to bridge metric spaces: Few-shot joint learning of intent detection and slot filling. *arXiv* **2021**, arXiv:2106.07343.
32. Liu, S.; Shi, Q.; Zhang, L. Few-shot hyperspectral image classification with unknown classes using multitask deep learning. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5085–5102. [CrossRef]
33. Zhang, L.; Yang, Q.; Liu, X.; Guan, H. Rethinking hard-parameter sharing in multi-domain learning. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 01–06. [CrossRef]
34. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
35. Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; Chi, E.H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In Proceedings of the 24th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 1930–1939. [CrossRef]
36. Tang, H.; Liu, J.; Zhao, M.; Gong, X. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In Proceedings of the 14th ACM Conference on Recommender Systems, Virtual Event, Brazil, 22–26 September 2020; pp. 269–278. [CrossRef]
37. Wang, Y.; Lam, H.T.; Wong, Y.; Liu, Z.; Zhao, X.; Wang, Y.; Chen, B.; Guo, H.; Tang, R. Multi-task deep recommender systems: A survey. *arXiv* **2023**, arXiv:2302.03525.
38. Pang, N.; Guo, S.; Yan, M.; Chan, C.A. A Short Video Classification Framework Based on Cross-Modal Fusion. *Sensors* **2023**, *23*, 8425. [CrossRef] [PubMed]
39. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.
42. Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. [CrossRef]
43. Wang, J.; Liu, H.; Ying, H.; Qiu, C.; Li, J.; Anwar, M.S. Attention-based neural network for end-to-end music separation. *CAAI Trans. Intell. Technol.* **2023**, *8*, 355–363. [CrossRef]
44. Yan, M.; Xiong, R.; Wang, Y.; Li, C. Edge Computing Task Offloading Optimization for a UAV-assisted Internet of Vehicles via Deep Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2023**, *73*, 5647–5658. [CrossRef]
45. Gu, Q.; Wang, Z.; Zhang, H.; Sui, S.; Wang, R. Aspect-Level Sentiment Analysis Based on Syntax-Aware and Graph Convolutional Networks. *Appl. Sci.* **2024**, *14*, 729. [CrossRef]
46. Ma, J.; Cheng, J.; Chen, Y.; Li, K.; Zhang, F.; Shang, Z. Multi-Head Self-Attention-Enhanced Prototype Network with Contrastive–Center Loss for Few-Shot Relation Extraction. *Appl. Sci.* **2024**, *14*, 103. [CrossRef]

47. Zhang, R.; Xue, C.; Qi, Q.; Lin, L.; Zhang, J.; Zhang, L. Bimodal Fusion Network with Multi-Head Attention for Multimodal Sentiment Analysis. *Appl. Sci.* **2023**, *13*, 1915. [CrossRef]
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
49. Guo, J.H.; Xu, N.N. Evaluation and cluster analysis of universities' transformation ability of scientific and technological achievements in China. *J. Intell.* **2016**, *35*, 155–168.
50. Liu, S.; Johns, E.; Davison, A.J. End-to-end multi-task learning with attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 1871–1880. [CrossRef]
51. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7482–7491.
52. Yang, Z.; Zhong, W.; Lv, Q.; Chen, C.Y.C. Multitask deep learning with dynamic task balancing for quantum mechanical properties prediction. *Phys. Chem. Chem. Phys.* **2022**, *24*, 5383–5393. [CrossRef] [PubMed]
53. Zhibin, S.; Yahong, Q.; Yu, G.; Hui, R. Research on emotion space of film and television scene images based on subjective perception. *J. China Univ. Posts Telecommun.* **2019**, *26*, 75–81. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.