IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 May 2025, accepted 29 July 2025, date of publication 6 August 2025, date of current version 12 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3596484

RESEARCH ARTICLE

The Effective Evaluation of Emotions in the Visual Emotion Images Using Convolutional Neural Networks

MODESTAS MOTIEJAUSKAS AND GINTAUTAS DZEMYDA
Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, 08412 Vilnius, Lithuania
Corresponding author: Modestas Motiejauskas (modestas.motiejauskas@mif.stud.vu.lt)

ABSTRACT This paper develops a model for recognizing emotions in visual images. The integration of contrastive-center loss optimization is proposed in this paper. This effectively improves the recognition of emotions when training a convolutional neural network against the baseline. The proposed contrastive-center loss function optimizes deep neural networks by enhancing feature discriminability. This loss function includes two key components: intra-class compactness and inter-class separability. We have suggested controlling the impact of the inter-class separability on the loss function. Moreover, we suggest combining cross-entropy and contrastive-center loss to calculate the total loss. In addition, we have proposed to apply the dimensionality reduction (visualization) for interactive evaluation of how the objects in the test set are arranged and how this arrangement, as well as the classification as a whole, can be improved by choosing the best combination of the strength of contrastive-center loss impact on the total loss. The efficiency of the developed model improvements is examined on three datasets: WEBEmo, FI-8, and EmoSet-118K. Our research allows us to improve the performance of visual emotion classification: for the WEBEmo dataset by 1.6%, the FI-8 dataset by 2.2%, and for the EmoSet-118K dataset by 2.52% higher accuracies than the baseline case.

INDEX TERMS Image analysis, convolutional neural network, CNN, contrastive-center loss, visual emotion images, evaluation of emotions, EfficientNet.

I. INTRODUCTION

Vision comprises the primary stimuli through which humans perceive external information. Images often convey emotional content that can evoke viewers' positive, negative, or neutral responses.

Visual emotion analysis aims to interpret these emotional cues in images. While traditional psychology identifies six basic emotions (happiness, anger, sadness, surprise, disgust, and fear), recent work by [1] and [2] extends this classification to seven categories: joy, sadness, surprise, disgust, anger, fear, and neutral. This field addresses three key analysis concepts: emotion recognition through facial expressions, general image emotion classification, and hybrid approaches combining both. Visual emotion analysis has been notable in recent years and has potential applications in various areas. Automated emotion recognition enables personalized human-computer interaction in mental health applications (for example, detecting depressive tendencies from social media images [3]) and adaptive education systems (for example, adjusting the pace of lessons based on student engagement detected in video feeds [4]). In robotics, emotion-aware systems improve social assistive devices for elderly care [5]. For the visual emotion analysis, psychologists primarily describe two types of emotion representation models or taxonomies. The first is categorical emotion states (CES), and the second is dimensional emotion space (DES) [6]. Plutchik [7] formulated a theoretical three-dimensional model of human emotions: a cone-shaped representation where the angle in the circle denotes the degree of similarity between emotions, and opposing emotions are expressed with vertical lines. Thayer [8] suggests defining

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

emotions according to arousal levels: energetic and tension. Emotional valence can be defined as either positive or negative - disgust is considered a negative emotion, whereas tenderness is the opposite. From the provided figure, it can be asserted that anger and tenderness have different emotional valences and are distinguishable in images.

It should be noted that people's subjectivity in identifying emotions in images is based on different methods. Therefore, differences between various methods, studies, or surveys in identifying emotions in images yield inconsistent results [7], [8]. In conclusion, there is no unified model for emotion research because human subjectivity, ambiguity, regional and cultural differences, and gender prevent this from being achieved [9], [10]. Our research gives the possibility to generalize people's emotional experiences using artificial intelligence.

One interesting application is emotion evaluation in artworks. The preliminary experiments are done in [11]. Here, the emotions in artworks by eight-year-old children and those of Vincent van Gogh were evaluated. The majority of predictions indicate sadness in Vincent van Gogh's paintings. For artworks by a child, there are primarily positive types of emotions.

Current visual emotion analysis identifies several research challenges. One such challenge is the affective emotion gap, which involves extracting features that can better distinguish closely related emotions [12]. The development of visual emotion analysis can be divided into several approaches: extraction of low-level visual features, analysis of mid-level visual features, and deep learning methods. These methods are frequently combined with various multi-scale models to address visual emotion prediction problems [13], [14], [15]. Luo et al. [13] propose a novel model for extracting emotion images' local and global features using visual transformers and a local attention module. Xu et al. [14] describe a multi-scale dependent attention network (MDAN), which leverages different emotion hierarchies. Zhang et al. [15] propose another network consisting of an affective region detection module and a multiscale feature module.

However, in these developments, the models are typically multi-stage and require complex feature extraction branches. Moreover, there is a lack of studies that bridge the affective emotion recognition gap; existing methods do not incorporate efficient techniques to improve the discriminability of image emotion classes.

To address these limitations, we propose a more robust approach that enforces better visual emotion class separability in the embedding space through contrastive-center loss optimization, thereby eliminating the need for complex multi-stage feature fusion and training. In this study, we analyze the high-dimensional feature outputs obtained from a convolutional neural network (CNN), a model that uses convolutional operations to extract hierarchical spatial features to derive robust feature representations for emotion recognition.

Our main contribution is an integration of contrastive-center loss optimization in the training of a deep neural network. We achieve several improvements for robust image emotion recognition by introducing contrastive-center loss optimization. Firstly, it brings images of the same emotion class closer in the embedding space. Secondly, it pushes different emotion classes further apart by leveraging class centers in the feature embeddings.

We investigate the trained network's emotion representations in the high-dimensional feature output space. We gain additional insights into the trained model by employing dimensionality reduction methods. In particular, we use the uniform manifold approximation and projection UMAP [16] dimension reduction technique to visualize the high-dimensional feature outputs. These visualizations allow us to understand positional embeddings, such as emotion groupings and overlapping emotion categories. Additionally, we conduct a cluster analysis of the high-dimensional feature outputs to establish the effectiveness of our proposed contrastive-center loss optimization.

The remainder of this paper is organized as follows: Section II reviews existing visual emotion recognition architectures and studies. Section III details our proposed contrastive-center loss integration. Section IV presents our experimental setup for training and evaluating described network. Section V presents comparative results between our improved model and the baseline approaches.

II. RELATED WORK

Zhao et al. [6] in their comprehensive survey review the advances in Affective Image Content Analysis (AICA) over the past twenty years. The review outlines the importance of understanding the emotional impact of images, which can convey rich semantics and evoke various emotions in viewers. The review covers key emotion representation models, available datasets, and state-of-the-art feature extraction and learning methods. It also discusses the applications of AICA in fields such as opinion mining, psychological health, business intelligence, and entertainment. The paper describes three main challenges in the affective image content analysis: affective gap, perception subjectivity, and label noise and absence. The affective gap refers to the difficulty in bridging the low-level features of images with the high-level emotional responses they evoke. For example, the authors illustrate the affective gap's significance by displaying two pictures with an ordinary physical object (e.g., a rose on a bright background or during the rain) – conveying different sentiments for the corresponding images.

Xu et al. [17] demonstrated a visual emotion recognition system that uses the CNN architecture. A CNN architecture-based model was trained to recognize objects, and then the problem was transferred to sentiment recognition. Chen et al. [18] used images labeled as Adjective-Noun Pairs ANPs. The authors managed to obtain statistical hints for emotion classification by manipulating the strength

of sentiment based on adjectives and nouns. However, these papers only demonstrate how to solve the binary emotion classification problems. You et al. [19] constructed a large-scale visual emotion dataset named Flickr and Instagram set (FI-8). This dataset was formulated according to the psychology studies and contains eight labeled emotion categories – amusement, awe, anger, contentment, disgust, excitement, fear, and sadness. This dataset was collected from freely available sources. Its images were manually labeled using the Amazon Mechanical Turk system, and the final Flickr and Instagram dataset has 23308 visual emotion images.

Other researchers analyse multi-layered network models to recognize and classify possible visual emotions [20]. These authors demonstrate the possibility of fusing visual semantic and visual-stream models for predicting emotions. Their proposed visual-semantic model produces possible visual-emotional embedding merging alongside the visual-stream model. Their Visual-semantic model is based on the DeepSentiBank structure [21], which produces conceptual emotion expression, e.g., a small beetle is expressed as the disgust expression. These expressions are formed as a graph embedding in a 2-dimensional space. They use the ResNet50 [22] model architecture for the visual stream emotion recognition model. The final fused model is the multiplication of these two different model architectures, and the visual emotion predictions are obtained in the result. A similar approach and study was done by Zhang et al. [23], where a multi-level representation model with side branches named Gram matrices for shallow features is proposed. The authors in [23] try integrating feature maps from different layers by applying a Gram matrix for further sentiment analysis – i.e., for negative and positive emotion classification. Based on the Gram matrix integration idea by Zhang et al. [23], authors of [24] have utilized Gram matrix modules for recognizing sadness emotion.

Xu et al. [25] introduce the multiple views prompt (MVP) model, which improves visual emotion recognition by integrating image content, generated captions, and enriched emotion labels through a structured prompting framework. The described method indicates an effective multi-modality feature fusion. Their proposed MVP method achieves state-of-the-art results on various visual emotion datasets. Luo et al. [26] describe a combined visual relationship feature and scene feature network CVRSF-Net – a dual-branch framework for image emotion recognition. Dual branches are defined as follows: the vision transformer encodes the entire image to a global feature map, and the visual-relationship feature branch highlights image emotion regions. This dual-branch network is fused using the graph attention network. In another study, Rui [27] extracts CNN features from each artwork image, embeds them in a low-dimensional space via a variational autoencoder, and then applies an unsupervised clustering algorithm (e.g., k-means) to classify the images into three sentiment groups – positive, negative, and neutral.

Sun et al. [28] propose a Supervised Contrastive Learning-based model for classifying image emotions. Their model integrates low-level handcrafted features (extracted using the LBP-U – Local Binary Patterns – algorithm [29]) and deep emotional features (learned through a ResNet-50 [22] encoder), combining them using a feature fusion strategy to enhance emotional classification performance. The authors also describe a novel two-stage training setup involving pre-training the ResNet-50 encoder using a supervised contrastive loss function to improve feature discrimination by reducing intra-class variability and enhancing inter-class separability. In the second stage, the pre-trained encoder is frozen, and the classifier is trained using cross-entropy loss optimization. Their findings indicate improvements over baseline methods on the FI image emotion dataset, suggesting the effectiveness of their approach. However, the study lacks the analysis of different emotion datasets to validate generalization. It does not extensively analyze the specific impact of the supervised contrastive learning optimization strategy on model performance.

The advancements of deep learning give a new option for image analysis from different viewpoints, including emotions. However, training deep neural networks needs many computing resources. The networks also tend to have vanishing or exploding gradient problems (see e.g. [30]). Batch normalization helps here, but the issues above remain with the increase in the model's depth. One solution was proposed in *Deep Residual Learning for Image Recognition* by [22] to use ResNet blocks, which connect the output of one layer with the input of an earlier layer. These skip connections are also commonly known as residual connections. Residual connections' applicability and usage have been proven widely in various architectures: Xception, MobileNetV2, DenseNet, EfficientNets [31], [32], [33], [34], [35]. Skip connections are also widely used in U-Net [36] and DeepLabV3 [37] for image segmentation tasks. Hybrid architecture-based models are emerging and achieving state-of-the-art results. For example, authors [38], [39] propose hybrid models that integrate convolutional neural networks (CNNs) with Vision Transformer (ViT) architectures. By combining CNNs with ViTs, these hybrid models leverage the strengths of both approaches – efficient local feature extraction from CNNs and the global attention capability of transformers. These architectures have demonstrated significant success across various fields, including medical imaging [40], remote sensing [41], [42], video surveillance [43], and anomaly detection [44].

III. IDEAS AND INSIGHTS FOR EFFECTIVE ANALYSIS OF VISUAL EMOTION IMAGES

In this section, we present our ideas and solutions for their integral use in emotion recognition in images of a general nature. Emotion analysis in this section mainly refers to the several steps and designs, which will be further described in the following subsections. We propose a model, analyze it, and identify and interpret its efficiency.

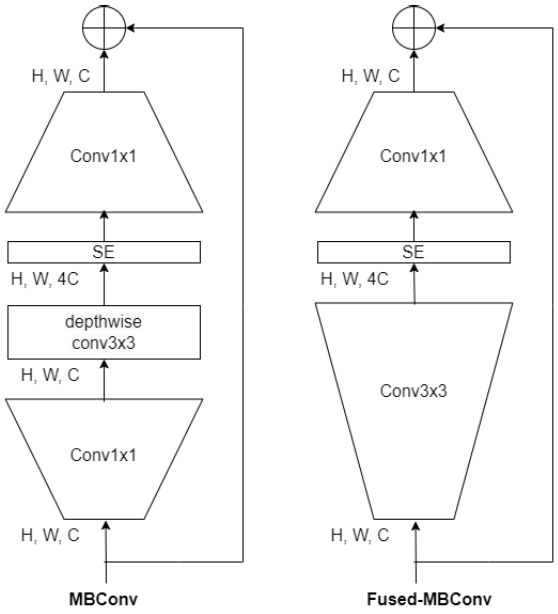


FIGURE 1. Structure of MBConv and Fused-MBConv blocks. Source: [35].

Challenges for integrating contrastive-center loss optimization are also described.

A. EfficientNetV2S CNN BACKBONE DESCRIPTION

This subsection describes key components of the EfficientNetV2S convolutional neural network [35]. We utilize the network as a feature extractor and the main component of our proposed model for image emotion recognition.

Figure 1 shows the structural blocks (modules) of the EfficientNetV2 architecture. The authors in [35] determined the entire network structure using combinations from these blocks. Here, H and W are the height and width of the input, and C is the number of channels. The MBConv block is an inverted residual block first introduced in the MobileNetV2 convolutional neural network architecture [32]. Initially, a 1×1 convolution expands the number of layer channels, followed by a special 3×3 depthwise convolution that reduces the number of parameters. Finally, a 1×1 convolution is applied to normalize the dimensions of the output and input. The authors of EfficientNetV2 [35] also improved this block with a so-called squeeze and excitation (SE) layer, which was first introduced in [45]. The essential difference between MBConv and Fused-MBConv is that Fused-MBConv replaces the first two layers with a conventional 3×3 convolution.

Table 1 illustrates the structure of EfficientNetV2S used in our paper. Stride refers to the convolution operation's step size. Channels No indicates the number of output channels from a particular block or operation. Layers No specifies the count of specific block repetitions within a certain stage. For example, the number of layers in the fourth stage, 6, indicates the number of MBConv block repetitions. MBConv[n] denotes the module MBConv with an expansion factor of n: the initial 1×1 convolution receives C channels and expands the output to $n \times C$ channels. SE0.25 refers to the squeeze and excitation block reduction ratio used to model channel-specific relations. In Stage 7, we have placed a global average pooling layer, resulting in a vector of 1792 fully connected units.

TABLE 1. Structure and parameters of EfficientNetV2S. MBConv and Fused-MBConv blocks are described in Figure 1. Source: [35].

Stage	Operation	Stride	Channels No.	Layers No.
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	256	15
7	Conv lxl & Pooling & FC	-	1792	1

B. THE PROPOSED NETWORK

Figure 2 shows the general scheme of our proposed model. It differs from that used in [11] and [24] because here we expand the number of features and use 136 feature vector layers instead of 128. The reason for such an extension is direct feature vector integration from both sides of the network. The general network is conceptually similar and based on the previously described architectures [23], [24]. For brevity, not all feature map outputs in Figure 2 have been directly named. In this case, those are named *blocks.1.2.add, blocks.2.3.add* and *blocks.4.8.add*. The backbone layers for feature extraction are not chosen arbitrarily or randomly. However, they closely resemble the feature extraction concept of popular backbones such as ResNet50-type networks. Figure 2 shows that the last layer feature map outputs are chosen from each stage. The output of the mentioned layers is passed to the corresponding Gram matrix module – each to its module. Each Gram matrix module output is concatenated, forming a vector of 136 features.

Figure 3 shows the Gram matrix module structure. In contrast with our paper [24], we reduced the final output dimensionality of the module, seeking to reduce the number of network parameters. Each module gets input, whose shape is $\mathbb{R}^{C \times H \times W}$, corresponding to the extracted layer's feature map, consisting of C feature sub-maps. Feature sub-maps are defined by height and width spatial dimensions $H \times W$. It should be noted that the output of the Gram matrix is in quadratic form and is expressed as $C \times C$ squared matrix. The Gram matrix is flattened into one 1-dimensional vector consisting of $C \times C$ units, which is further compressed by a dense layer resulting in $C/2$ units, which are then applied by an activation function and batch. Here, $C/2$ means dividing the number of obtained original channels by a factor of 2. SiLU activation function is called the sigmoid linear unit [46], or more commonly known as a swish activation function. Accordingly, the other side of the Gram matrix module consists of a 1×1 convolution operation.

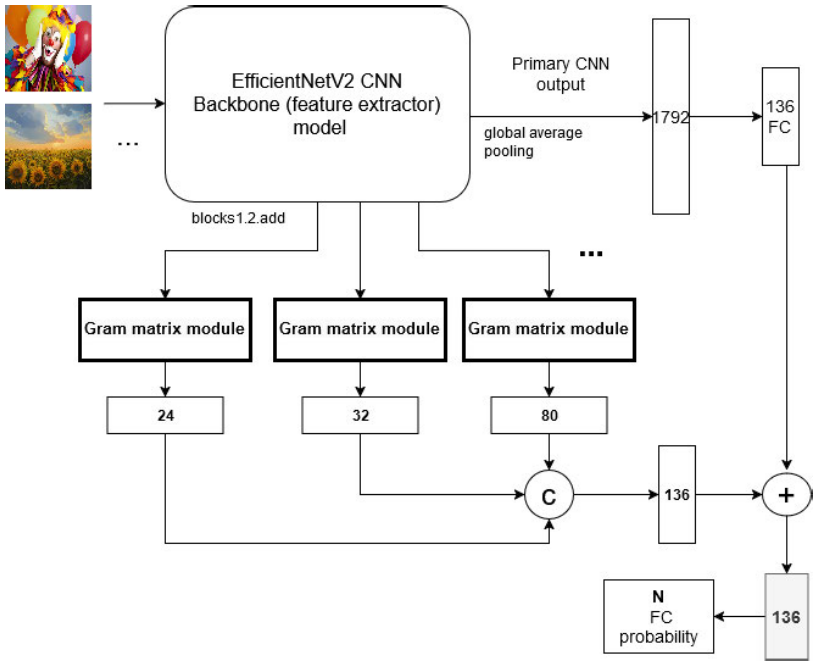


FIGURE 2. General schema of the proposed network, used in this paper.

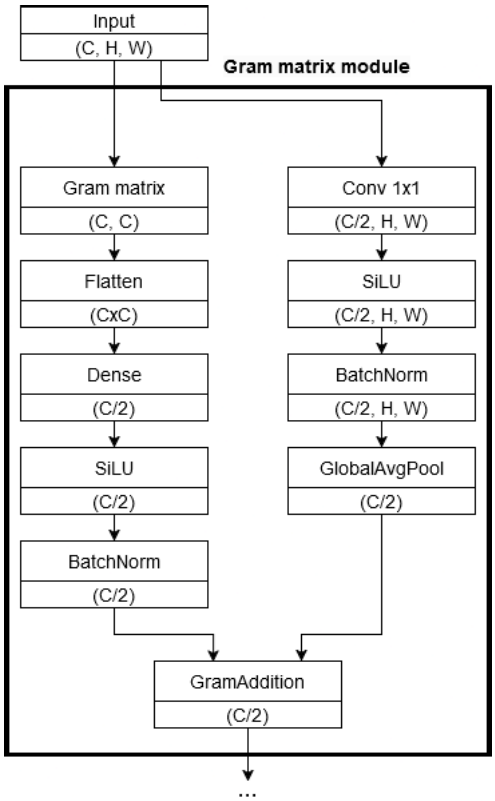


FIGURE 3. Proposed in [24] Gram matrix module schema with detailed flow.

From this convolution, we get $C/2$ feature sub-maps. The corresponding SiLU activation function is applied – for each feature sub-map among $C/2$ ones, a singular average value is computed from all $H \times W$ values of the sub-map. As a result, we obtain a vector of $C/2$ length that contains average values of all $C/2$ sub-maps. The final result of the Gram matrix module is fused by feature-wise addition from each side of the branch as shown in Figure 2. We also considered the concatenation, multiplication, and average fusion strategies, but those options increased the number of trainable parameters and gave us no gains.

C. CONTRASTIVE-CENTER LOSS

Our model has a penultimate layer output in Figure 2. The network's second-to-last layer is the penultimate layer, highlighted in greyscale in Figure 2. The output of this layer is a vector of 136 elements – features. This feature vector is applied to compute the contrastive-center loss. The feature vector $\mathbf{x}_k$ of sample k will comprise 136 features. We also have C number of class centers $\mathbf{c}_{y_i}$ corresponding to the class label y_i. Each class center corresponds to a different emotion.

The class centers $\mathbf{c}_i$ are learnable parameters initialized randomly and updated via gradient descent alongside the model parameters during training. Each class center $\mathbf{c}_i$ is a 136-dimensional vector, matching the dimensionality of the feature vector $\mathbf{x}_k$. Contrastive-center loss inclusion for the described model (see Figure 2) is achieved through an auxiliary module.

The contrastive-center loss [47] is a loss function designed to optimize deep neural networks by enhancing feature discriminability. The contrastive-center loss is defined as follows:

$$L_{\text{contr}} = L_{\text{intra}} + L_{\text{inter}}. \tag{1}$$

This loss function includes two key components: intra-class compactness and inter-class separability. Intra-class

compactness L_{intra} aims to minimize the distances between feature embeddings and their corresponding class centers. In other words, this part seeks to bring feature vectors corresponding to the proper class center closer. The second term, which is inter-class loss L_{inter}, enforces a margin (m) between the centers of different classes to maximize their separation.

The intra-class compactness component is as follows:

$$L_{\text{intra}} = \frac{1}{N}\sum_{k=1}^{N} \|\mathbf{x}_k - \mathbf{c}_{y_k}\|^2, \tag{2}$$

where $\mathbf{x}_k$ is a feature vector of the k-th sample (image), $\mathbf{c}_{y_k}$ is a center of class, to which the k-th sample belongs. The total number of classes is C. Denote the centers of classes by $\mathbf{c}_1, \ldots, \mathbf{c}_C$. Each center of the class is a learnable vector, initialized randomly and updated via gradient descent during training. The centers are model parameters (like weights in a neural network) optimized alongside the rest of the network during training.

This loss component minimizes the distance between sample feature vectors and their corresponding class centers. The distance is computed as the squared Euclidean norm, $\|\mathbf{x}_k - \mathbf{c}_{y_k}\|^2$, which enforces tighter clustering of feature vectors by corresponding class.

The inter-class separability loss component is defined as follows:

$$L_{\text{inter}} = \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j\neq i} \max(0, m - \|\mathbf{c}_i - \mathbf{c}_j\|)^2, \tag{3}$$

where C is the number of classes, $\mathbf{c}_i$ and $\mathbf{c}_j$ are the centers of the i-th and j-th classes, respectively, m is the margin that forces a minimum separation between class centers, $\|\mathbf{c}_i - \mathbf{c}_j\|$ is the Euclidean distance between the centers of classes i and j. $\max(0, \cdot)$ indicates no penalty is applied if the distance between centers exceeds the specified margin. This loss component aims to improve inter-class separability by penalizing pairs of class centers closer than the specified margin m, pushing them farther apart in the embedding space. So in this case, the proper margin m value selection is a crucial parameter, which can be considered a hyperparameter requiring additional tuning.

We suggest extending the loss function:

$$L_{\text{contr}} = L_{\text{intra}} + \lambda \cdot L_{\text{inter}}, \tag{4}$$

where parameter λ allows to control the strength of separability among inter-class centers.

The hyperparameters λ and m require dataset-specific tuning, which we suggest be performed during training. Contrastive-center loss optimization is an auxiliary term, and in the next section, we describe its integration for our image emotion recognition problem.

The selection of λ (which controls inter-class separability in L_{inter}) was tested empirically using the FI-8 dataset. We observed weak performance for $\lambda < 1$ compared to $\lambda = 1$, while $\lambda > 5$ led to overfitting. Since the optimal λ depends on the specific dataset, we chose $\lambda = 1$ as a simple, stable value that avoids these extremes.

D. INTEGRATING CONTRASTIVE-CENTER LOSS FOR IMAGE EMOTION RECOGNITION

We suggest integrating the contrastive-center loss into the training process. It may also be used for testing purposes, such as evaluating class centers and computing their inter-distances. It has been noted in the previous studies [6] that some image emotion categories tend to be closely related (or overlapping). Integrating the contrastive center-loss into our model can address the observed gap. This subsection describes the strategy and proposition behind contrastive-center loss integration for image emotion recognition.

Another objective function for CNN training is called categorical cross-entropy loss [48]. It is the typical objective loss function upon various CNNs being trained. This cross-entropy loss function is expressed as below:

$$L_{\text{entr}} = -\frac{1}{N}\sum_{k=1}^{N} \log(p_{y_k}), \tag{5}$$

where L_{entr} is the average loss for the entire subset (training or validation). In our case, the number of classes C is the number of emotion categories. N corresponds to the number of training or validation samples, depending on the phase of the learning process. p_{y_k} is the predicted probability of the true class for the k-th sample, y_k is the label of the true class for the k-th sample.

We can gather feature outputs simultaneously from the main and penultimate layers. The main layer is fully connected with C feature outputs corresponding to the class probabilities. We can compute both losses (cross-entropy and contrastive-center) and combine them into total loss L_{total}:

$$L_{\text{total}} = L_{\text{entr}} + \beta \cdot L_{\text{contr}}, \tag{6}$$

where in Eq. 6, L_{entr} is the computed cross-entropy loss from the main layer, and L_{contr} is the contrastive-center loss from the penultimate layer. The weight coefficient β allows us to control the strength of the contrastive-center loss impact. The coefficient $\beta = 0$ is for the case of training without contrastive-center loss integration (see Figure 2).

E. RELATION BETWEEN CLUSTERS OF FEATURE VECTORS AND CLASSIFICATION

One of the primary goals of our study is to investigate the trained model, whose structure is described in Figure 2. By encouraging intra-class compactness and inter-class separation, the contrastive-center loss is designed to enhance cluster cohesion and separation. We evaluate whether this proposed improvement aligns with empirical clustering quality and classification accuracy. To achieve this, we analyze emotion representations in the high-dimensional feature

FIGURE 4. Emotion examples in the CAER-S dataset.

space of the trained network and suggest the following possibilities:

- Visualizing feature embeddings in 2D space using dimensionality reduction methods, e.g., UMAP [16].
- Cluster analysis, e.g., using k-means, and evaluation of clustering quality by using metrics such as adjusted Rand index (ARI) [49], normalized mutual information (NMI) [49], and ambiguous sample ratio (ASR) [50].

Visualization and cluster analysis are applied to the set of feature vectors obtained from the penultimate layer when the test dataset is shown to the trained network. In both cases, the high-dimensional feature outputs of the same penultimate layer serve as data for analysis. This enables the evaluation of the performance of CNN on the test dataset, both visually and from the point of view of similarities/dissimilarities of classes (in our case – emotions). Different clustering quality metrics allow us to evaluate the obtained clusters from various standpoints.

The adjusted Rand index (ARI) quantifies the similarity between the predicted clusters and the ground-truth labels by adjusting for chance agreement. Similarly, the normalized mutual information (NMI) measures the mutual dependence between two clusterings, with normalization ensuring that the values lie within a consistent range. Both metrics are robust to imbalanced class distributions.

The ambiguous sample ratio ASR can be defined in the following way. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of N testing samples, and let there be C clusters with known centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_C$. Assume $d(\cdot, \cdot)$ is a chosen distance metric, and $\delta > 0$ is a constant predefined distance threshold.

We define the *ambiguity* of a single sample $\mathbf{x}_k$ as

$$\text{Ambiguity}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } \sum_{j=1}^{C} \mathbf{1}\left[\, d(\mathbf{x}_k, \mathbf{c}_j) < \delta \,\right] > 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Here, $\mathbf{1}[\cdot]$ is the indicator function, equal to 1 if the condition inside is true, and 0 otherwise.

Finally, the *ambiguous sample ratio* (ASR) is defined as

$$\text{ASR} = \frac{1}{N} \sum_{k=1}^{N} \text{Ambiguity}(\mathbf{x}_k). \tag{8}$$

A sample $\mathbf{x}_k$ is considered ambiguous if it is close – within distance threshold δ – to more than one cluster center. ASR is the ratio of all samples that meet the chosen ambiguity criterion. The main problem is the selection of the distance threshold δ. We suggest computing the pairwise distances of the points from X and selecting the threshold δ, the largest distance among the 10 percent smallest distances. This criterion is heuristic and lacks formal justification, but it is an option for evaluation purposes. This approach closely resembles methods in fuzzy clustering or probabilistic models (e.g., Gaussian Mixture Models), which quantify ambiguity using continuous membership values [50], [51], [52].

In the experiments below, we utilized k-means clustering and set the number of clusters the same as the number of prediction classes from the corresponding dataset.

IV. EXPERIMENTAL SETUP

This subsection describes the datasets, the methodology, the pre-processing steps, the training strategy, the objective function optimization, and the metrics used to evaluate results.

The primary dataset for our study is the subset of WEBEmo for binary classification problems. The dataset is described in [2] and how the subset has been gathered in [24]. Like in [24], a 61074 filtered images dataset has been obtained, where about 46% of images represent sadness emotion. This dataset has been divided into 80% training, 10% validation, and 10% testing subsets. The WEBEmo training subset contains 26445 images expressing no sadness and 22413 images conveying sadness. Similarly, the WEBEmo validation subset consists of 3284 images expressing no sadness and 2823 images expressing sadness. Finally, the testing subset split is divided into the same ratio as the validation subset, consisting of the same number of images in each class.

To assess the network generalization capabilities, the Flickr and Instagram (FI-8) dataset [19] was used, too. It comprises 23308 images labeled according to Mikels' emotion hierarchy [53].

Further, we evaluated the network on EmoSet [54], a large-scale visual emotion dataset designed for Visual Emotion Analysis (VEA). EmoSet includes two subsets:

- EmoSet-3.3M: Contains 3.3 million images retrieved and annotated using automated methods.
- EmoSet-118K: A human manually labeled subset of 118102 images, where each image is labeled according to one of eight emotion categories based on Mikels' model [53].

TABLE 2. Emotion distribution in EmoSet-118K and FI-8 datasets.

Dataset	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sadness	Total
EmoSet-118K	19445	10660	15037	16337	10666	19828	13453	12676	118102
FI-8	4924	1266	3151	5374	1658	2963	1032	2922	23308

Table 2 shows the emotion category distribution in EmoSet-118K and FI-8.

This indicates a relatively even and balanced distribution across positive (amusement, awe, contentment, excitement) and negative (anger, disgust, fear, sadness) emotions, making the dataset EmoSet-118K highly suitable for further evaluation studies. The most common categories are amusement and contentment, whilst the least popular would be anger and disgust.

We also employed the CAER-S (Context-Aware Emotion Recognition - Static) dataset [55], a subset of the larger CAER dataset. CAER-S comprises training and testing sets without a validation subset and includes emotion categories: anger, disgust, fear, happy, neutral, sad, and surprise. Notably, the training set contains exactly 7001 images per category, and the testing set includes 2999 images per category, indicating a perfectly balanced distribution. Figure 4 displays sample images from this dataset. Authors of CAER-S [55] note that the emotion images were gathered from 79 TV shows.

We trained four model cases, each corresponding to one of the four visual emotion datasets. The dataset-dependent hyperparameters are the learning rate, number of training epochs, β, margin m, and δ. We train with stochastic gradient descent (SGD) with momentum 0.9 and an initial learning rate of 0.02. We run for 20 epochs (50 on the CAER-S dataset) with a batch size of 256. The learning rate is then adjusted using a Cosine Annealing with Warm Restarts schedule [56]: it decays from 0.02 down to 1×10^{-4} and restarts back every 5 epochs. The parameter β allows us to influence the effectiveness of contrastive-center loss. Margin m was initially set to 1.5, but later larger distances were also considered. The parameter λ was set to 1. During training, we first apply a random resized crop to 224×224 and a random horizontal flip. We then use the RandAugment augmentation technique [57] with distortion strength set to 3, and number of transformations set to 2.

The integration of the contrastive-center loss (refer to subsection III-D) can lead to exploding gradients. Gradient is a vector that indicates the direction and magnitude in which the neural network's parameters should be adjusted to reduce the training error [48]. To mitigate this, we employed gradient clipping (normalization). Specifically, given the original gradient vector g, the clipped gradient g_{clipped} is computed as:

$$g_{\text{clipped}} = g \cdot \min\left(1, \frac{h}{\|g\|_2}\right), \tag{9}$$

where h is the threshold for the maximum allowed gradient norm. If $\|g\|_2 > h$, the gradient is scaled to have an L2 norm equal to h, reducing the risk of numerical instability. In our experiments, we set $h = 5$.

V. RESULTS

A. EXPERIMENTS: COMPARISON OF METRICS

The experimental study aims to investigate the emotion representations by the trained network. Utilizing the high-dimensional feature vectors, we aim to evaluate the effectiveness of contrastive-center loss integration using previously defined metrics. We chose several hyperparameter values seeking to analyze a wider definition area.

TABLE 3. Performance metrics on dependence on β; WEBEmo sadness testing set.

β	Accuracy	ARI	NMI	ASR
0.0	0.8214	0.0987	0.1884	0.8797
0.1	0.8253	0.1309	0.2126	0.7888
0.3	0.8274	0.1222	0.2147	0.7318
0.4	0.8278	0.1221	0.2155	0.6703
0.5	0.8281	0.1255	0.2153	0.6082
0.8	0.8266	0.1384	0.2185	0.4095
1.0	0.8287	0.1371	0.2177	0.3505

Table 3 presents the performance of the model on the WEBEmo sadness testing set. Here, the hyperparameter β controls the penalization strength of the contrastive-center loss (refer to subsection III-D). Case $\beta = 0$ corresponds to the baseline model. Each row reports the performance of a trained model on a binary emotion classification task. Accuracy denotes the ratio of correctly predicted samples. The highest accuracy is achieved at $\beta = 1.0$, indicating a 0.7% improvement over the baseline. The main increase in accuracy is achieved when the weight coefficient β increases from 0 to 0.3. Additionally, ARI, NMI, and ASR indicate that the clustering quality of the penultimate layer's feature representations is enhanced when the contrastive-center loss is integrated.

TABLE 4. Performance metrics on dependence on β; FI-8 testing set.

β	Accuracy	ARI	NMI	ASR
0	0.6968	0.3907	0.4337	0.8262
0.1	0.7132	0.4735	0.4706	0.2659
0.3	0.7124	0.4498	0.4650	0.2492
0.4	0.7138	0.4486	0.4648	0.2172
0.5	0.7153	0.4565	0.4692	0.1858
0.8	0.7135	0.4574	0.4688	0.1341
1.0	0.7127	0.4233	0.4597	0.2433

In Table 4, the performance results are gathered from evaluating the model on the FI-8 testing set. In our case,

the network has been trained to recognize emotion from 8 classes. Accuracy indicates the ratio of correctly predicted image emotion samples. Highest accuracy was obtained with $\beta = 0.5$. From the established baseline, we achieved 1.8% improvement of accuracy. The highest ARI, NMI scores were achieved with $\beta = 0.1$, and lowest ASR score was achieved with $\beta = 0.8$. However, accuracy improvement is marginal. In our case, we can note that clustering quality improves when integrating the contrastive-center loss optimization. This suggests the idea that the model gains the capability to recognize emotion classes better.

TABLE 5. Performance metrics on dependence on β; EmoSet-118K testing set.

β	Accuracy	ARI	NMI	ASR
0	0.7794	0.3416	0.4605	0.8156
0.1	0.7858	0.5400	0.5833	0.3766
0.3	0.7859	0.5565	0.5945	0.2560
0.4	0.7870	0.5608	0.5983	0.2257
0.5	0.7866	0.5630	0.6008	0.1995
0.8	0.7879	0.5618	0.6026	0.1517
1.0	0.7880	0.5656	0.6055	0.1298

Table 5 summarizes the performance on the EmoSet-118K testing set for an 8-class emotion recognition problem. The highest accuracy is achieved at $\beta = 1.0$ (an improvement of approximately 0.9% over the baseline). The best clustering quality, reflected by the highest ARI and NMI scores, is observed at $\beta = 1.0$, although the differences are marginal. The lowest ASR is observed at $\beta = 1.0$. Overall, these results confirm that integrating the contrastive-center loss enhances feature clustering. This indicates that the trained network performs better in tasks with more emotion classes.

TABLE 6. Performance metrics on dependence on β; CAER-S testing set.

β	Accuracy	ARI	NMI	ASR
0.0	0.9104	0.7139	0.7438	0.7963
0.1	0.9106	0.8100	0.7880	0.0547
0.3	0.9118	0.8142	0.7919	0.0227
0.4	0.9114	0.7103	0.7425	0.0083
0.5	0.9106	0.7289	0.7671	0.0041
0.8	0.9112	0.7082	0.7533	0.0034
1.0	0.9120	0.7295	0.7670	0.0007

In Table 6, performance results are gathered from evaluating the CAER-S testing set. The results are obtained from the trained model using the described CAER-S dataset [55]. The network has been trained to recognize emotion from 7 classes. The highest accuracy was obtained for $\beta = 1.0$. Compared to the baseline, integrating contrastive-center loss optimization did not result in a substantial improvement in accuracy. The highest ARI and NMI scores were achieved with $\beta = 0.3$. However, differences are marginal.

In Tables 7, 8, and 9 performance results are gathered evaluating the parameter margin m influence. We can observe that a larger margin improves classification accuracy, up to some point. Margin selection to $m = 3$ or $m = 5$ produces the best overall performances with marginal differences.

TABLE 7. Performance metrics depending on margin; WEBEmo sadness testing set.

Margin	Accuracy	ARI	NMI	ASR
1	0.8312	0.1335	0.2216	0.5016
1.5	0.8281	0.1331	0.2203	0.3628
3	0.8330	0.1584	0.2242	0.0851
5	0.8374	0.1689	0.2293	0.0665
10	0.8366	0.1862	0.2323	0.0219

TABLE 8. Performance metrics depending on margin; EmoSet-118K testing set.

Margin	Accuracy	ARI	NMI	ASR
1	0.7997	0.5886	0.6179	0.0966
1.5	0.7992	0.5871	0.6182	0.0918
3	0.8033	0.5877	0.6230	0.0508
5	0.8025	0.5872	0.6223	0.0191
10	0.8012	0.5133	0.5938	0.0455

TABLE 9. Performance metrics on dependence on margin; FI-8 testing set.

Margin	Accuracy	ARI	NMI	Ambiguous ratio
1	0.7112	0.4644	0.4682	0.1908
1.5	0.7138	0.4650	0.4693	0.1832
3	0.7171	0.4657	0.4701	0.1121
5	0.7150	0.4725	0.4685	0.0296
10	0.7138	0.4832	0.4795	0.0082

B. EXPERIMENTS: VISUAL ANALYSIS

Human emotions often overlap because emotions are complex, interconnected, and influenced by multiple factors at once. Here are some key reasons why emotional overlap occurs. Different emotions can trigger similar physical reactions. Situations often evoke more than one emotion. Emotions aren't binary, they lie on a continuum. As emotions shift gradually, their boundaries can blur. The way we interpret events can cause different emotional blends. Cultural norms or personal experiences shape how emotions are processed and expressed, often blending one feeling with another. In this section, we also see such overlaps of emotions visually. However, we do not try to explain the reasons for the overlaps.

Let us consider the visualization of the results of the second-to-last layer. Here, we have the high-dimensional feature vector for each test image. The total number of such vectors equals the number of images in the set of test images. Let us use, e.g., the UMAP method [16] for the dimensionality reduction and visualization of the set of such vectors. It is not the only possible visualization method for this purpose. The initial dimensionality of the feature vector is 136.

In the first experiment, the network is trained on the WEBEmo dataset [24]. The results are given in Figure 5.

(a) Contrastive-center loss integration term $\beta = 0$.

(b) Contrastive-center loss integration term $\beta = 0.5$

FIGURE 5. Visualization of the trained model results; WEBEmo.

Two models were used: (a) baseline, which has a multiplier $\beta = 0$ at the integrated contrastive-center loss in the total loss L_{total}, and (b) $\beta = 0.5$. In both cases, we have visualized 6108 vectors, corresponding to the 6108 emotion images in the testing subset. A single point in the figure is a representation on a plane of a particular point from a test set, and the coloring of the point is of ground-truth labels (true class).

We can observe no clear, distinct boundary between the two emotion groups. However, we observe a polarity (contrast) of the distribution of points on the plane: emotion classes tend to form apparent groups – see a distribution of colors in Figure 5. An exciting discovery is that in case (b), the distribution of points on a plane has some more regular form, and the polarity (separation) becomes clearer. However, overlap is still apparent in (b) - we might expect an alignment of 82.81% classification accuracy.

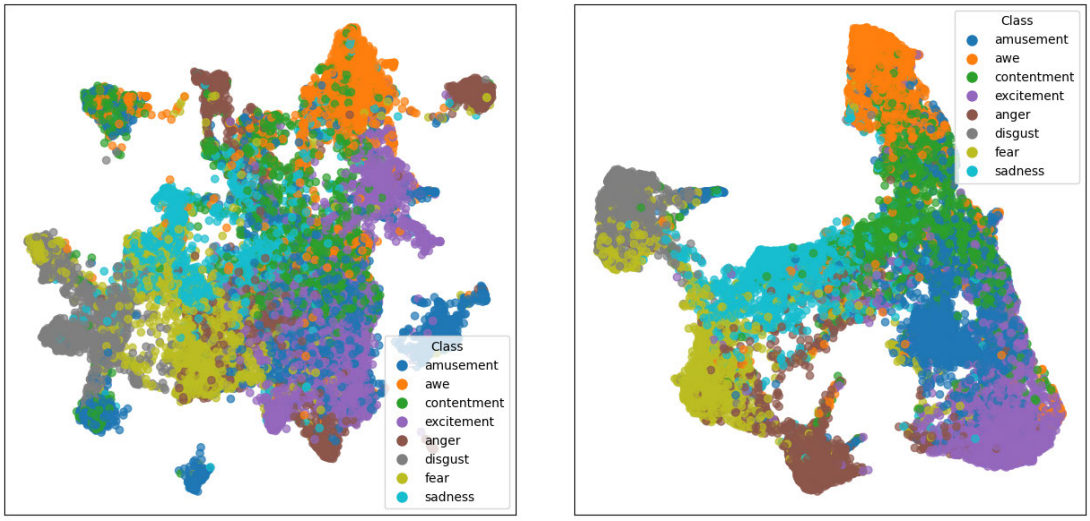
In the second experiment, the network is trained on the EmoSet-118K dataset [54]. The results are given in Figure 6. Two models were used: (a) baseline, $\beta = 0$, and (b) integrated contrastive-center loss optimization, $\beta = 0.5$. In both cases, we visualized 17716 vectors, corresponding to the 17716 emotion images in the testing subset. A single point in the figure is a representation on a plane of a particular point from a test set, and the coloring of the point is of ground-truth labels. Total number of classes is 8: anger, disgust, fear, sadness, amusement, awe, contentment, and excitement. We observe clusters of emotions in (a). In addition to the clusters, we see some positive and negative emotions polarization in (b). Therefore, a subtle separation exists between the negative and positive emotion groupings in the plot. Class overlap is apparent in (a) and (b).

The network is trained on the FI-8 [19] in the third experiment. The results are given in Figure 7. Two models were used: (a) baseline, $\beta = 0$, and (b) integrated contrastive-center loss optimization, $\beta = 0.5$. In both cases, we visualized 3407 vectors, corresponding to the 3407 emotion images in the testing subset. A single point in the figure is a representation on a plane of a particular point from a test set, and the coloring of the point is of ground-truth labels. The total number of classes is 8: anger, disgust, fear, sadness, amusement, awe, contentment, and excitement. We observe clusters of emotions in (a). In addition to the clusters, we see some positive and negative emotions polarization in (b). Therefore, there is a subtle separation between the plot's negative and positive emotion groupings. Class overlap is apparent in (a) and (b). Nevertheless, the contrastive-center loss integration effect appears to be effective. We can also note that visualized points for fear and anger emotions appear not to have tight groups and are highly overlapped. There are about 160 samples each for these two emotions. Compared to the other classes, there are 400-750 samples for each emotion group, meaning the trained network still has difficulty identifying minority-class emotions.

The experiments in this section lead to the idea that we can visually evaluate the quality of network training. A more concentrated distribution of points inside their clusters corresponding to the particular emotions means a better classification. We see this when $\beta = 0.5$ in the total loss L_{total}.

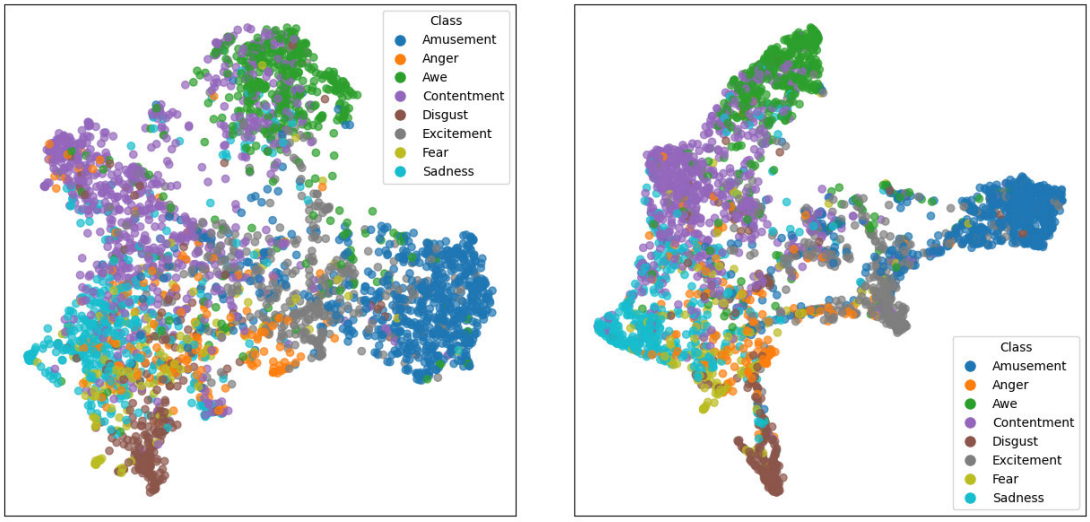
C. EXPERIMENTS: DISTRIBUTION OF THE CLASS CENTERS

Let us consider a trained network. In Section III-C, we have introduced the term of class center. The class centers $\mathbf{c}_{y_i}$ are learnable parameters initialized randomly and updated via gradient descent alongside the model parameters during training.



(a) $\beta = 0$ (b) $\beta = 0.5$

FIGURE 6. Visualization of the trained model results; EmoSet-118K.



(a) $\beta = 0$ (b) $\beta = 0.5$

FIGURE 7. Visualization of the trained model results; dataset FI-8.

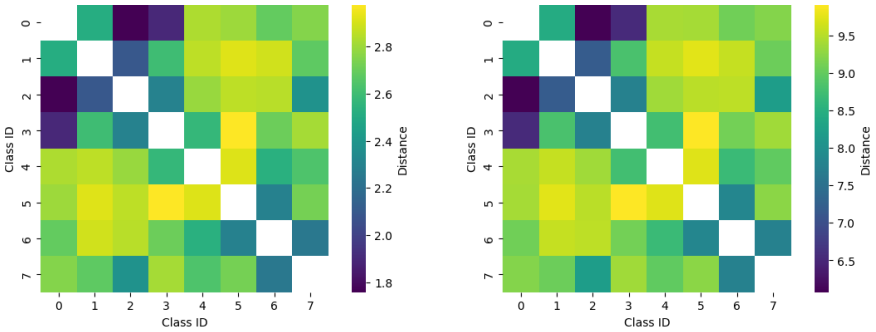
In particular, we are concerned primarily with the inter-class separability loss component L_{inter}. This component has two parts. Firstly, centers that correspond to the emotion class. Secondly, margin m that enforces minimum distances between pair-wise centers. This means that during the training, every class center is pushed further away from every other class center by a given margin. Therefore, the information on how well the trained network responds to the specified margin values would be helpful in the additional evaluation of training quality.

In Figures 8 and 9, the pair-wise center distance matrices are displayed for two data sets, EmoSet-118K and FI-8, in the case of different margins $m = 3$, $m = 5$, and $m = 10$.

Both plots in Figure 8 show that the trained model sufficiently maintains the selected distances with different margins. Several emotion class centers are too close, and this leads to the model gaps in discerning those emotions.

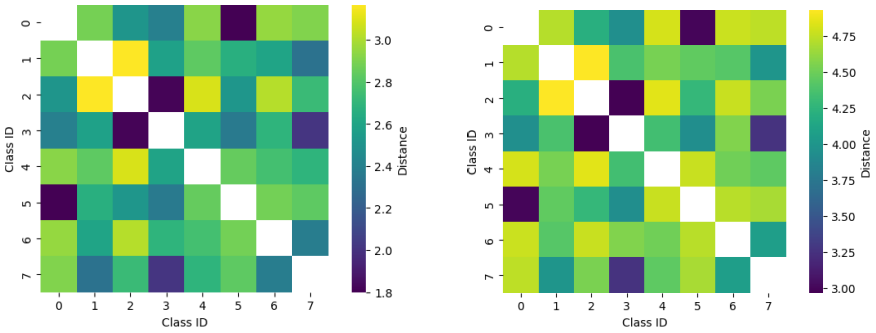
Figure 9 indicates that training the network using a larger margin $m = 5$ as compared to $m = 3$ yields a slightly higher mean distance. It is also evident that on both plots in Figure 9 the chosen margin is maintained relatively well.

Experiments indicate that the margin parameter m significantly influences classification performance. Using the baseline model, accuracies of 82.14% on the WEBEmo dataset, 77.94 % on the EmoSet-118K dataset, and 69.68% on the FI-8 dataset are achieved. Through empirical evaluation,



(a) margin $m = 3$ (b) margin $m = 10$

FIGURE 8. Trained model pair-wise center distance matrix. The model was trained on the EmoSet-118K dataset.



(a) margin $m = 3$ (b) margin $m = 5$

FIGURE 9. Trained model pair-wise center distance matrix. The model was trained on the FI-8 dataset.

the optimal margin value was determined to be $m = 5$, which generalized well across all three datasets (WEBEmo, FI-8, and EmoSet-118K). With this margin, the final improved model achieves 83.74% accuracy on WEBEmo, 71.88% accuracy on FI-8, and 80.46% accuracy on EmoSet-118K. Therefore, we see a good responsiveness of the training process to the selected hyperparameter margin m.

TABLE 10. Comparison to baselines across three test-sets.

WEBEmo sadness	Accuracy	ARI	NMI	ASR
Baseline	0.8214	0.0987	0.1884	0.8797
Ours	0.8374	0.1689	0.2293	0.0665
EmoSet-118K	Accuracy	ARI	NMI	ASR
Baseline	0.7794	0.3320	0.4550	0.3567
Ours	0.8046	0.5988	0.6241	0
FI-8	Accuracy	ARI	NMI	ASR
Baseline	0.6968	0.3907	0.4337	0.8262
Ours	0.7188	0.4592	0.4628	0.0646

Another hyperparameter for weighting inter-class separability loss component L_{inter} is λ (refer Eq. 4). The best accuracies were obtained with the following λ values: for FI-8 case $\lambda = 100$, for EmoSet-118K case $\lambda = 100$, and WEBEmo case $\lambda = 5$. We see a large diversity in optimal λ values for different data sets. However, the influence of λ is insignificant when its values grow: it suffices to use λ more or less 5.

Finally, we can conclude from the experiments (see aggregated results in Table 10) that integrating the contrastive-center loss optimization is beneficial – the model manages to differentiate between emotion centers.

VI. CONCLUSION

The paper addresses a significant and growing area in affective computing – emotion recognition from visual imagery using Convolutional Neural Networks (CNNs). This area is particularly relevant in human-computer interaction, marketing, psychology, and multimedia analysis. If properly executed, using CNNs to evaluate visual emotion images can significantly enhance automatic emotion recognition capabilities, thereby contributing to the existing literature.

This paper develops the model for recognizing visual image emotions, see Figure 2. The development is taking place in the integration of contrastive-center loss optimization. Such integration effectively improves the recognition of emotions when training deep neural networks against the baseline.

The contrastive-center loss is a function designed to optimize deep neural networks by enhancing feature discriminability. This loss function includes two key components: intra-class compactness and inter-class separability. We have

suggested utilizing the weight coefficient to control the impact of the inter-class separability in the loss function. Moreover, we suggest combining cross-entropy and contrastive-center losses into the total loss. The additional weight coefficient is introduced to control the strength of the contrastive-center loss's impact on the total loss.

Visualization is to see how the objects (images) in the test set are arranged in the 2D plane and how this arrangement, as well as the classification as a whole, can be improved by choosing the best combination of the strength of contrastive-center loss impact on the total loss.

Inter-class separability is essential for the classification of images, depending on their emotions. We showed the possibility of controlling the inter-class separability using the margin hyperparameter that forces a minimum separation between class centers. We discover that the margin values set to $margin = 3$ or $margin = 5$ are good initial choices.

The efficiency of the developed model is examined on three datasets: WEBEmo, FI-8, and EmoSet-118K. Our proposed methods showed us visual emotion classification performance improvements: for the WEBEmo dataset by 1.6%, the FI-8 dataset by 2.2%, and for the EmoSet-118K dataset by 2.52% higher accuracies.

New avenues could be explored for integrating contrastive-center loss when training a deep neural network. More effective contrastive-center loss definitions might exist, without tuning (requirement of) related hyperparameters. Furthermore, several gaps need to be addressed or considered: inherent emotion ambiguity and noisy emotion image datasets. Emotions are interpreted subjectively. Therefore, our proposed method relies heavily on the quality of visual emotion datasets. Integration of contrastive-center loss requires additional tuning of hyperparameters, which can limit applicability. The next gap that might arise is from the weak discriminative power of the backbone.

The proposed model for evaluating emotions in visual emotion images will be a valuable tool for psychologists, art critics, designers, architects, and human-computer interaction developers. Such computer-based emotion recognition will also be used by artificial intelligence. This opens up a wide range of opportunities for applied research.

ACKNOWLEDGMENT

The authors are thankful to the Information Technology Research Center of Vilnius University for the high-performance computing resources.

REFERENCES

[1] R. Karbauskaitė, L. Sakalauskas, and G. Dzemyda, "Kriging predictor for facial emotion recognition using numerical proximities of human emotions," *Informatica*, vol. 31, pp. 249–275, Feb. 2020.

[2] R. Panda, J. Zhang, H. Li, J. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *Proc. Comput. Vis.-ECCV*, 2018, pp. 594–612.

[3] A. G. Reece and C. M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Sci.*, vol. 6, no. 1, Dec. 2017, Art. no. 15.

[4] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.

[5] H. Abdollahi, M. H. Mahoor, R. Zandie, J. Siewierski, and S. H. Qualls, "Artificial emotional intelligence in socially assistive robots for older adults: A pilot study," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2020–2032, Jul. 2023.

[6] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: Two decades review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6729–6751, Oct. 2022.

[7] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *Amer. Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[8] R. E. Thayer, *The Biopsychology of Mood and Arousal*. London, U.K.: Oxford Univ. Press, 1990.

[9] R. Arya, J. Singh, and A. Kumar, "A survey of multidisciplinary domains contributing to affective computing," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100399.

[10] K. Cortiñas-Lorenzo and G. Lacey, "Toward explainable affective computing: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 13101–13121, Oct. 2024.

[11] M. Motiejauskas and G. Dzemyda, "Evaluation of emotions in artworks using EfficientNet convolutional network integrating the Gram matrix modules," in *Proc. 4th Int. Conf. Artif. Intell., Robot., Commun. (ICAIRC)*, Dec. 2024, pp. 882–887.

[12] G. Zhao, H. Yang, B. Tu, and L. Zhang, "A survey on image emotion recognition," *J. Inf. Process. Syst.*, vol. 17, no. 6, pp. 1138–1156, 2021.

[13] Y. Luo, X. Zhong, M. Zeng, J. Xie, S. Wang, and G. Liu, "CGLF-net: Image emotion recognition network by combining global self-attention features and local multiscale features," *IEEE Trans. Multimedia*, vol. 26, pp. 1894–1908, 2024.

[14] L. Xu, Z. Wang, B. Wu, and S. Lui, "MDAN: Multi-level dependent attention network for visual emotion analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9469–9478.

[15] H. Zhang and M. Xu, "Multiscale emotion representation learning for affective image recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 2203–2212, 2023.

[16] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[17] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," *Neurocomputing*, vol. 291, pp. 187–194, May 2018.

[18] E. Dellandrea, N. Liu, and L. Chen, "Classification of affective semantics in images based on discrete and dimensional models of emotions," in *Proc. Int. Workshop Content Based Multimedia Indexing (CBMI)*, Jun. 2010, pp. 1–6.

[19] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," 2016, *arXiv:1605.02677*.

[20] H. Yang, Y. Fan, G. Lv, S. Liu, and Z. Guo, "Exploiting emotional concepts for image emotion recognition," *Vis. Comput.*, vol. 39, no. 5, pp. 2177–2190, May 2023.

[21] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks," 2014, *arXiv:1410.8586*.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[23] H. Zhang, Y. Liu, D. Xu, K. He, G. Peng, Y. Yue, and R. Liu, "Learning multi-level representations for image emotion recognition in the deep convolutional network," *Proc. SPIE*, vol. 12083, pp. 636–646, Feb. 2022.

[24] M. Motiejauskas and G. Dzemyda, "EfficientNet convolutional neural network with Gram matrices modules for predicting sadness emotion," *Int. J. Comput. Commun. CONTROL*, vol. 19, no. 5, p. 6697, Sep. 2024.

[25] Q. Xu, Y. Wei, S. Yuan, J. Wu, L. Wang, and C. Wu, "Learning emotional prompt features with multiple views for visual emotion analysis," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102366.

[26] Y. Luo, X. Zhong, J. Xie, and G. Liu, "CVRSF-net: Image emotion recognition by combining visual relationship features and scene features," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 9, no. 3, pp. 2321–2333, Jun. 2025.

[27] X. Rui, "A convolutional neural networks based approach for clustering of emotional elements in art design," *PeerJ Comput. Sci.*, vol. 9, p. e1548, Sep. 2023.

[28] J. Sun, Q. Zhang, K. Yuan, Y. Jiang, and X. Chen, "A supervised contrastive learning-based model for image emotion classification," *World Wide Web*, vol. 27, no. 3, May 2024, Art. no. 29.

[29] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," in *Proc. 12th Int. Conf. Pattern Recognit.*, vol. 1, 1994, pp. 582–585.

[30] L. Borawar and R. Kaur, "ResNet: Solving vanishing gradient in deep networks," in *Proc. Int. Conf. Recent Trends Comput.*, 2023, pp. 235–247.

[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.195

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[33] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[34] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[35] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 10096–10106.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[37] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[38] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[39] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5260–5269.

[40] P. Dutta, K. A. Sathi, M. A. Hossain, and M. A. A. Dewan, "Conv-ViT: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection," *J. Imag.*, vol. 9, no. 7, p. 140, Jul. 2023. [Online]. Available: https://www.mdpi.com/2313-433X/9/7/140

[41] G. Wang, H. Chen, L. Chen, Y. Zhuang, S. Zhang, T. Zhang, H. Dong, and P. Gao, "P2FEViT: Plug-and-play CNN feature embedded hybrid vision transformer for remote sensing image classification," *Remote Sens.*, vol. 15, no. 7, p. 1773, Mar. 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/7/1773

[42] C. Zhang, W. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. Wang, "Transformer and CNN hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022.

[43] W. Ullah, T. Hussain, F. U. M. Ullah, M. Y. Lee, and S. W. Baik, "TransCNN: Hybrid CNN and transformer mechanism for surveillance anomaly detection," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106173. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623003573

[44] Q. Zhu, Q. Wu, Y. Yue, Y. Bao, T. Zhang, X. Wang, Z. Jiang, and H. Chen, "Vision transformer–based anomaly detection method for offshore platform monitoring data," *Struct. Control Health Monitor.*, vol. 2024, no. 1, Jan. 2024, Art. no. 1887212. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2024/1887212

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2018, pp. 7132–7141.

[46] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018.

[47] C. Qi and F. Su, "Contrastive-center loss for deep neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2851–2855. [Online]. Available: http://ieeexplore.ieee.org/document/8296803/

[48] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[49] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Clustering validity checking methods: Part II," *ACM SIGMOD Rec.*, vol. 31, no. 3, pp. 19–27, Sep. 2002.

[50] E. S. Dalmaijer, C. L. Nord, and D. E. Astle, "Statistical power for cluster analysis," *BMC Bioinf.*, vol. 23, no. 1, Dec. 2022, Art. no. 205.

[51] M. Wegmann, D. Zipperling, J. Hillenbrand, and J. Fleischer, "A review of systematic selection of clustering algorithms and their evaluation," 2021, *arXiv:2106.12792*.

[52] M. Ulu and Y. S. Türkan, "Cluster analysis and comparative study of different clustering performance and validity indices," in *Proc. Int. Symp. Prod. Res.*, 2024, pp. 33–45.

[53] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behav. Res. Methods*, vol. 37, no. 4, pp. 626–630, Nov. 2005. [Online]. Available: http://link.springer.com/10.3758/BF03192732

[54] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, "EmoSet: A large-scale visual emotion dataset with rich attributes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20326–20337.

[55] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10142–10151.

[56] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–11.

[57] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.

MODESTAS MOTIEJAUSKAS received the B.S. degree in information technologies and the M.S. degree in computer modeling from Vilnius University, in 2018 and 2021, respectively, where he is currently pursuing the Ph.D. degree with the Institute of Data Science and Digital Technologies. His current research interests include deep neural networks, computer vision, and visual image emotion analysis.

GINTAUTAS DZEMYDA received the Doctoral (Ph.D.) degree in technical sciences and the Doctor Habilis degree from Kaunas University of Technology, in 1984 and 1997, respectively. He was conferred the title of a Professor with Kaunas University of Technology and Vilnius University. Currently, he is a Professor, a Principal Researcher, and the Head of the Cognitive Computing Group of Institute of Data Science and Digital Technologies, Vilnius University. He has extensive experience in research at the intersection of computer science and artificial intelligence with other urgent disciplines, including medicine, economics, and technology. He is the author of more than 290 scientific publications, two monographs, and 17 edited books. He is a keynote speaker at 14 international conferences. His research interests include artificial intelligence, data mining, multidimensional data visualization, optimization theory and applications, neural networks, and image analysis. He was a Full Member of the Lithuanian Academy of Sciences, in 2011. He was awarded the Lithuanian Science Prize twice, in 2001 and 2021. He is the Editor-in-Chief of two international journals *Informatica* and *Baltic Journal of Modern Computing*.